

# Digital Design & Computer Arch.

## Lecture 1: Introduction and Basics

Prof. Onur Mutlu

ETH Zürich

Spring 2020

20 February 2020

# Brief Self Introduction

---



## ■ Onur Mutlu

- ❑ Full Professor @ ETH Zurich CS (EE), since September 2015
- ❑ Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-...
- ❑ PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
- ❑ <https://people.inf.ethz.ch/omutlu/>
- ❑ [omutlu@gmail.com](mailto:omutlu@gmail.com) (Best way to reach me)
- ❑ <https://people.inf.ethz.ch/omutlu/projects.htm>

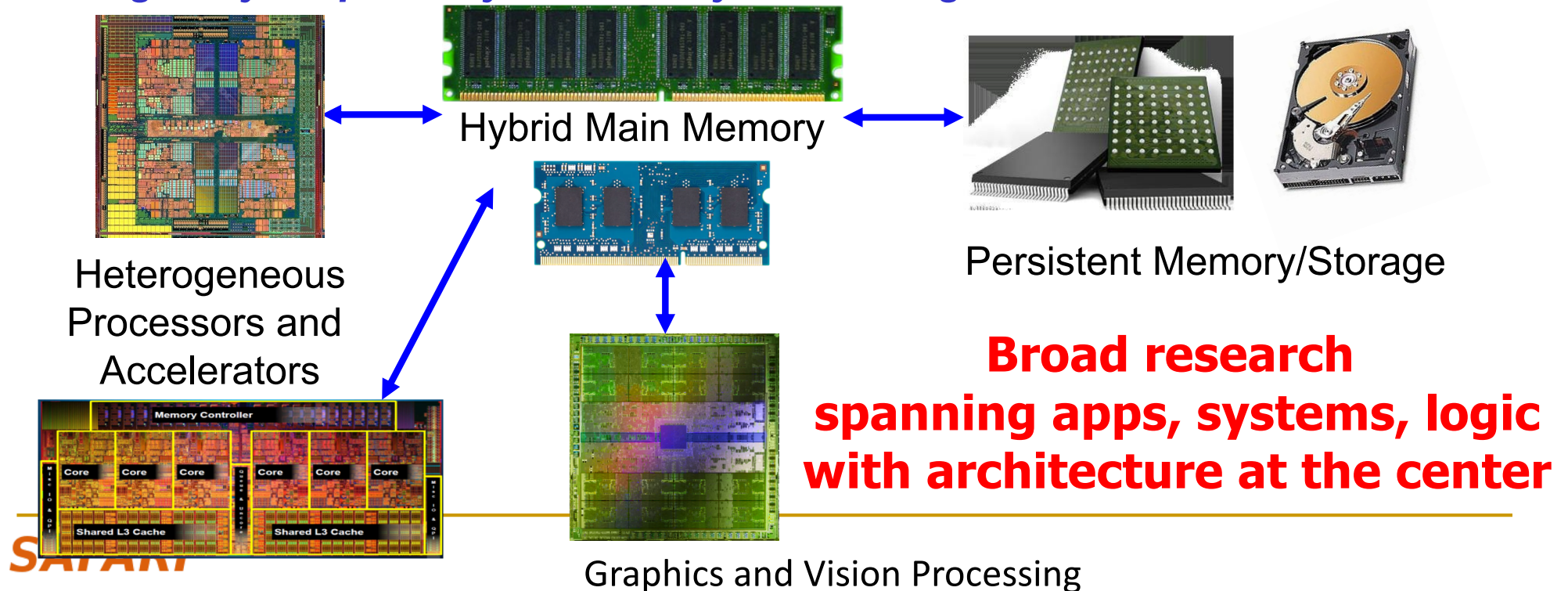
## ■ Research and Teaching in:

- ❑ Computer architecture, computer systems, hardware security, bioinformatics
- ❑ Memory and storage systems
- ❑ Hardware security, safety, predictability
- ❑ Fault tolerance
- ❑ Hardware/software cooperation
- ❑ Architectures for bioinformatics, health, medicine
- ❑ ...

# Current Research Focus Areas

**Research Focus: Computer architecture, HW/SW, bioinformatics, security**

- **Memory and storage (DRAM, flash, emerging), interconnects**
- **Heterogeneous & parallel systems, GPUs, systems for data analytics**
- **System/architecture interaction, new execution models, new interfaces**
- **Hardware security, energy efficiency, fault tolerance, performance**
- **Genome sequence analysis & assembly algorithms and architectures**
- **Biologically inspired systems & system design for bio/medicine**



# Four Key Directions

---

- Fundamentally **Secure/Reliable/Safe** Architectures
- Fundamentally **Energy-Efficient** Architectures
  - **Memory-centric** (Data-centric) Architectures
- Fundamentally **Low-Latency and Predictable** Architectures
- Architectures for **AI/ML, Genomics, Medicine, Health**

# What Will We Learn in This Course?

# How Computers Work

(from the ground up)

**And Why We Care**

# Why Do We Have Computers?

# Why Do We Do Computing?

# To Solve Problems

# To Gain Insight

To Enable  
a Better Life & Future

# How Does a Computer Solve Problems?

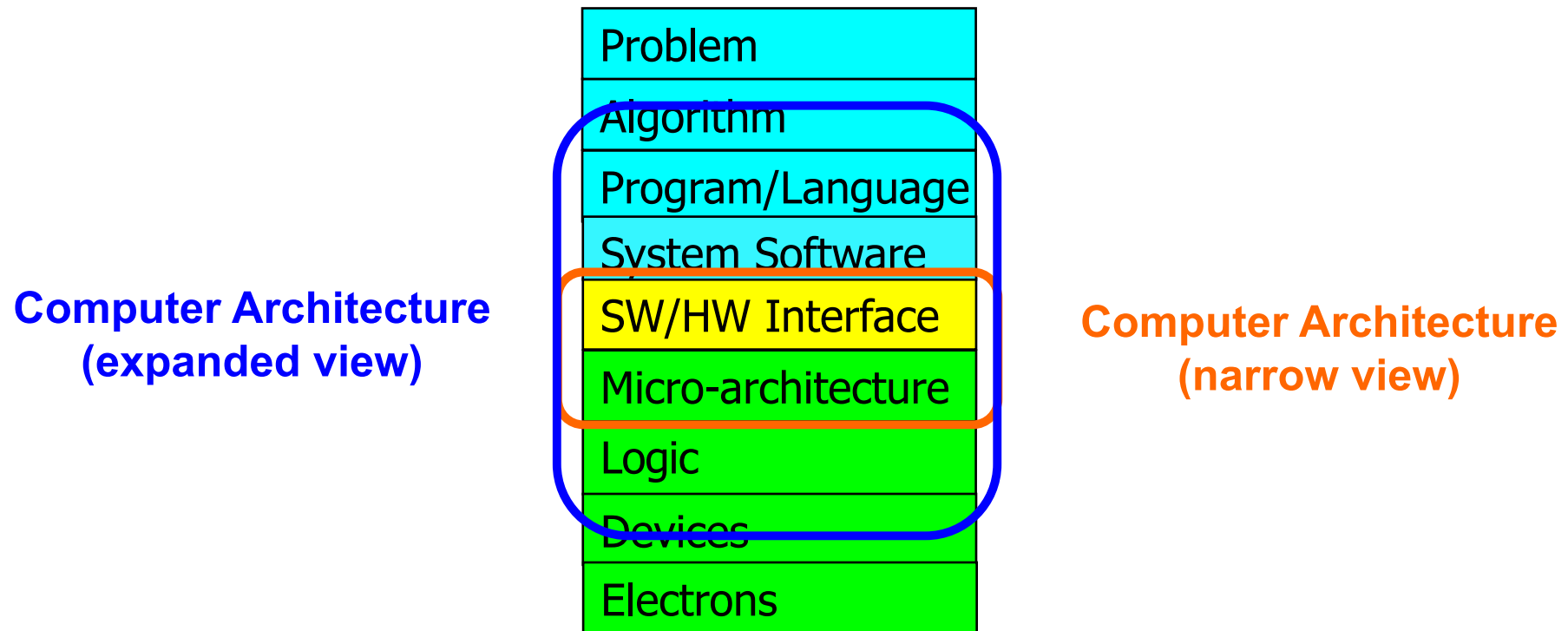
# Orchestrating Electrons

In today's dominant technologies

# How Do Problems Get Solved by Electrons?

# The Transformation Hierarchy

---



# Levels of Transformation

“The purpose of computing is [to gain] insight” (*Richard Hamming*)  
*We gain and generate insight by solving problems*  
*How do we ensure problems are solved by electrons?*

## Algorithm

Step-by-step procedure that is **guaranteed to terminate** where **each step is precisely stated** and **can be carried out by a computer**

- **Finiteness**
- **Definiteness**
- **Effective computability**

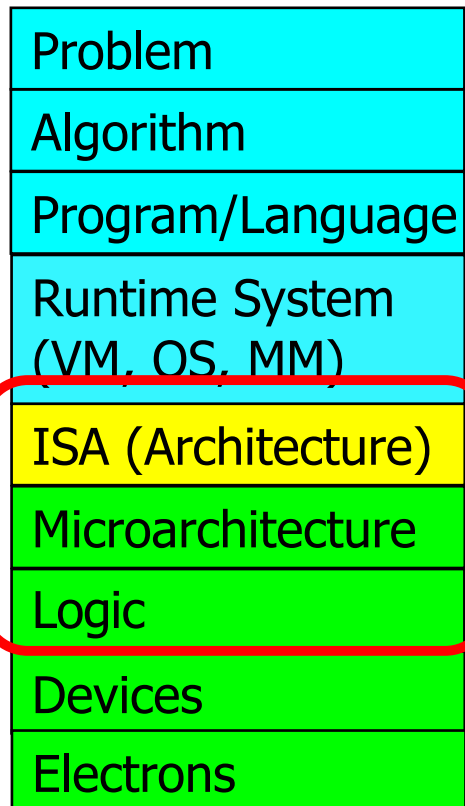
Many algorithms for the same problem

## Microarchitecture

An implementation of the ISA

## Digital logic circuits

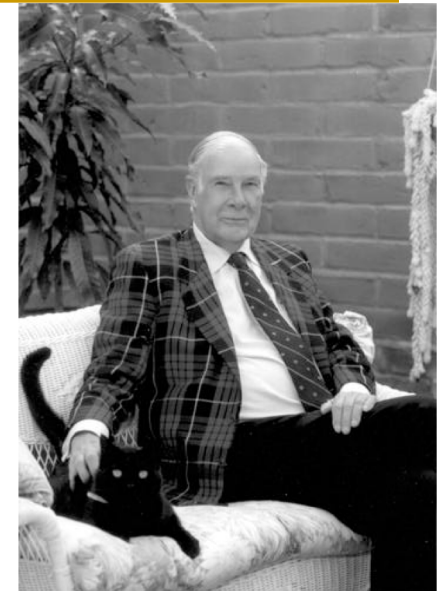
Building blocks of micro-arch (e.g., gates)



ISA  
(Instruction Set Architecture)

Interface/contract between  
SW and HW.

What the programmer  
assumes hardware will  
satisfy.



# Computer Architecture

---

- is the **science** and **art** of designing **computing platforms** (hardware, interface, system SW, and programming model)
- to achieve a set of **design goals**
  - E.g., highest performance on earth on workloads X, Y, Z
  - E.g., longest battery life at a form factor that fits in your pocket with cost < \$\$\$ CHF
  - E.g., best average performance across all known workloads at the best performance/cost ratio
  - ...
- Designing a supercomputer is different from designing a smartphone → But, many fundamental principles are similar

# Different Platforms, Different Goals

---



# Different Platforms, Different Goals



# Different Platforms, Different Goals

---



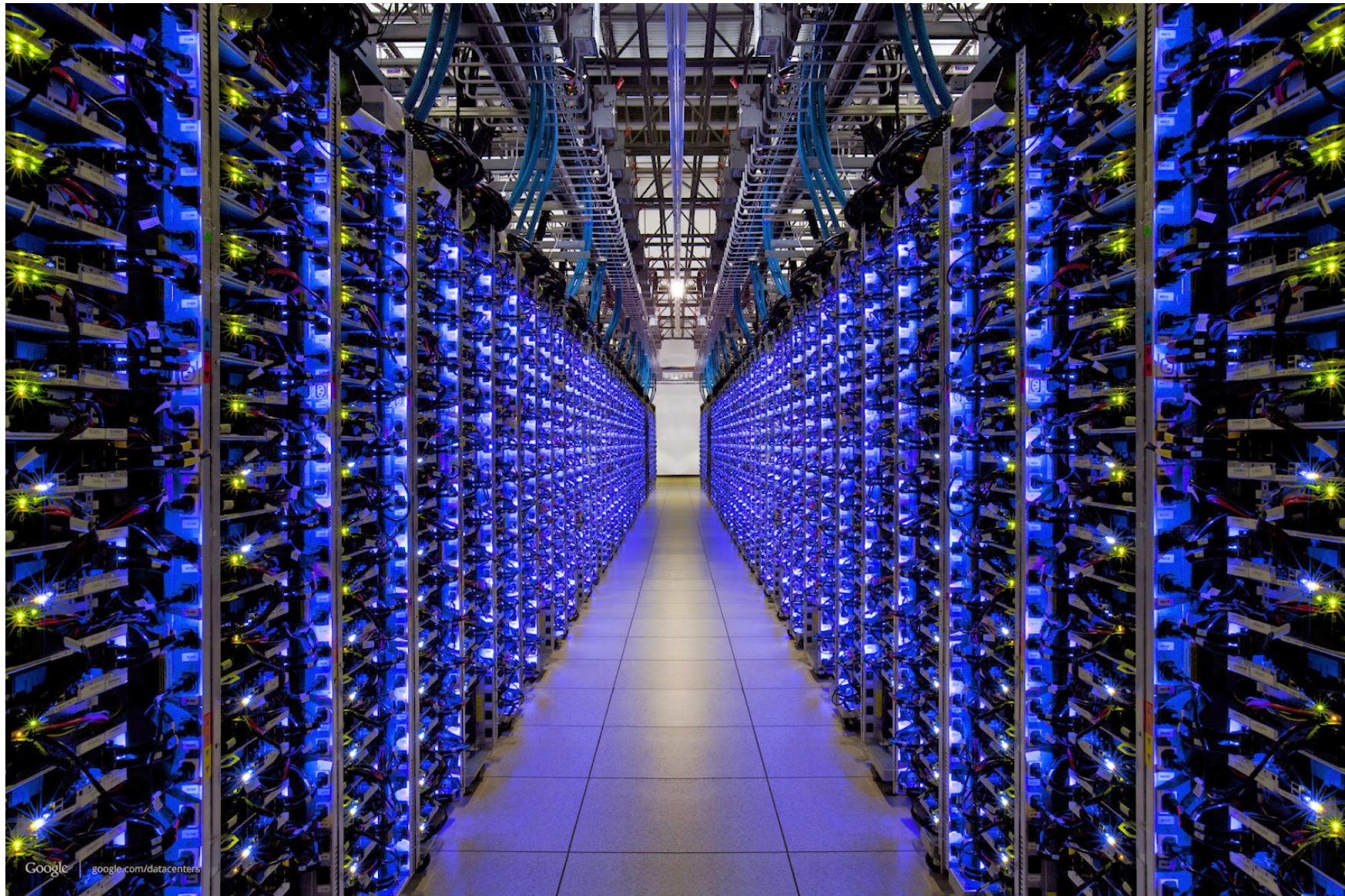
# Different Platforms, Different Goals

---



# Different Platforms, Different Goals

---

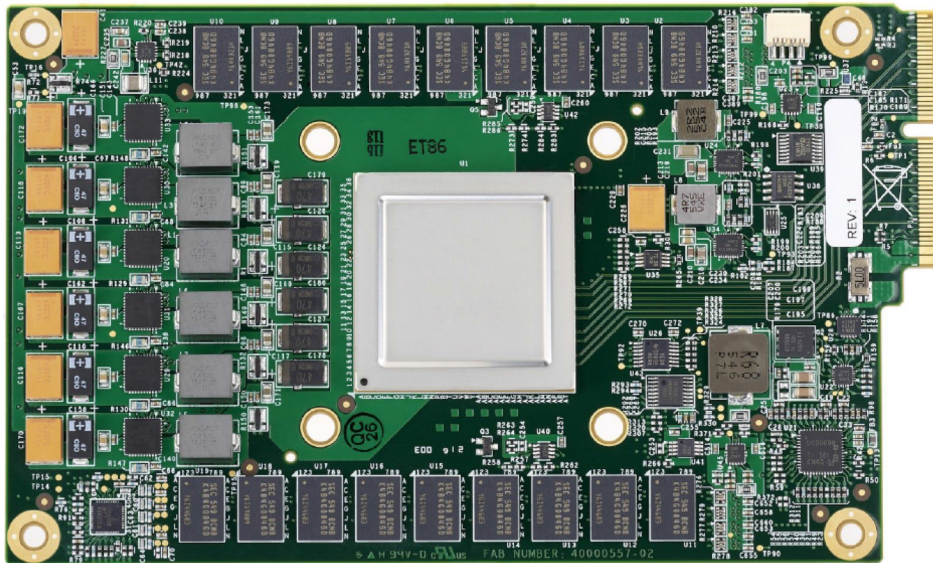


# Different Platforms, Different Goals

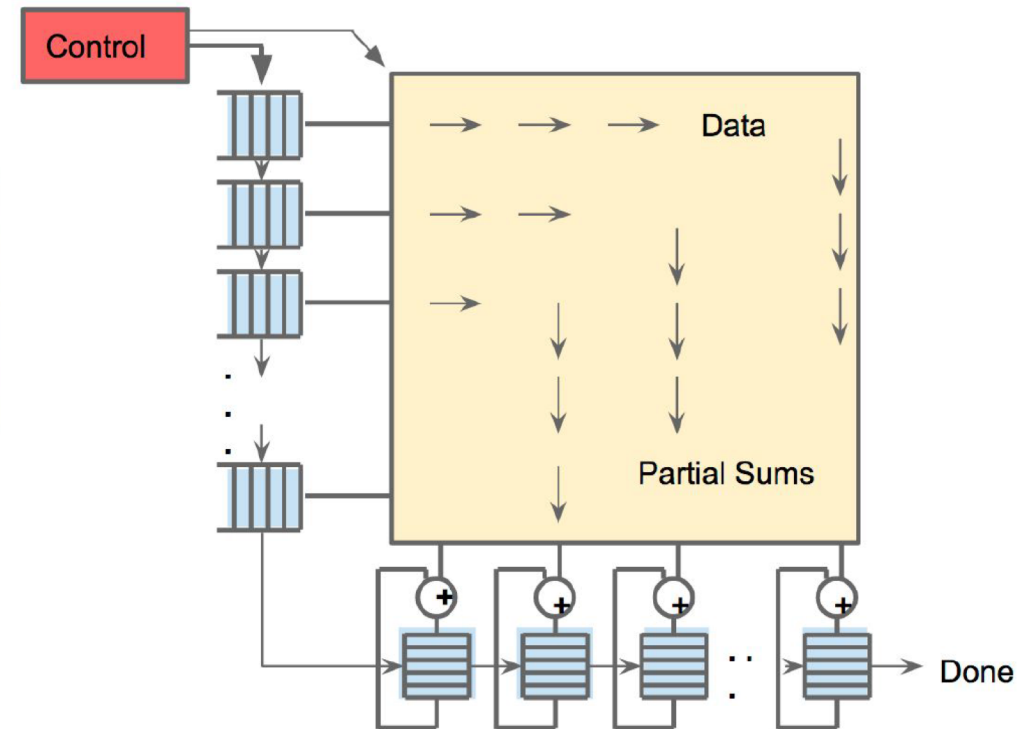
---



# Different Platforms, Different Goals



**Figure 3.** TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

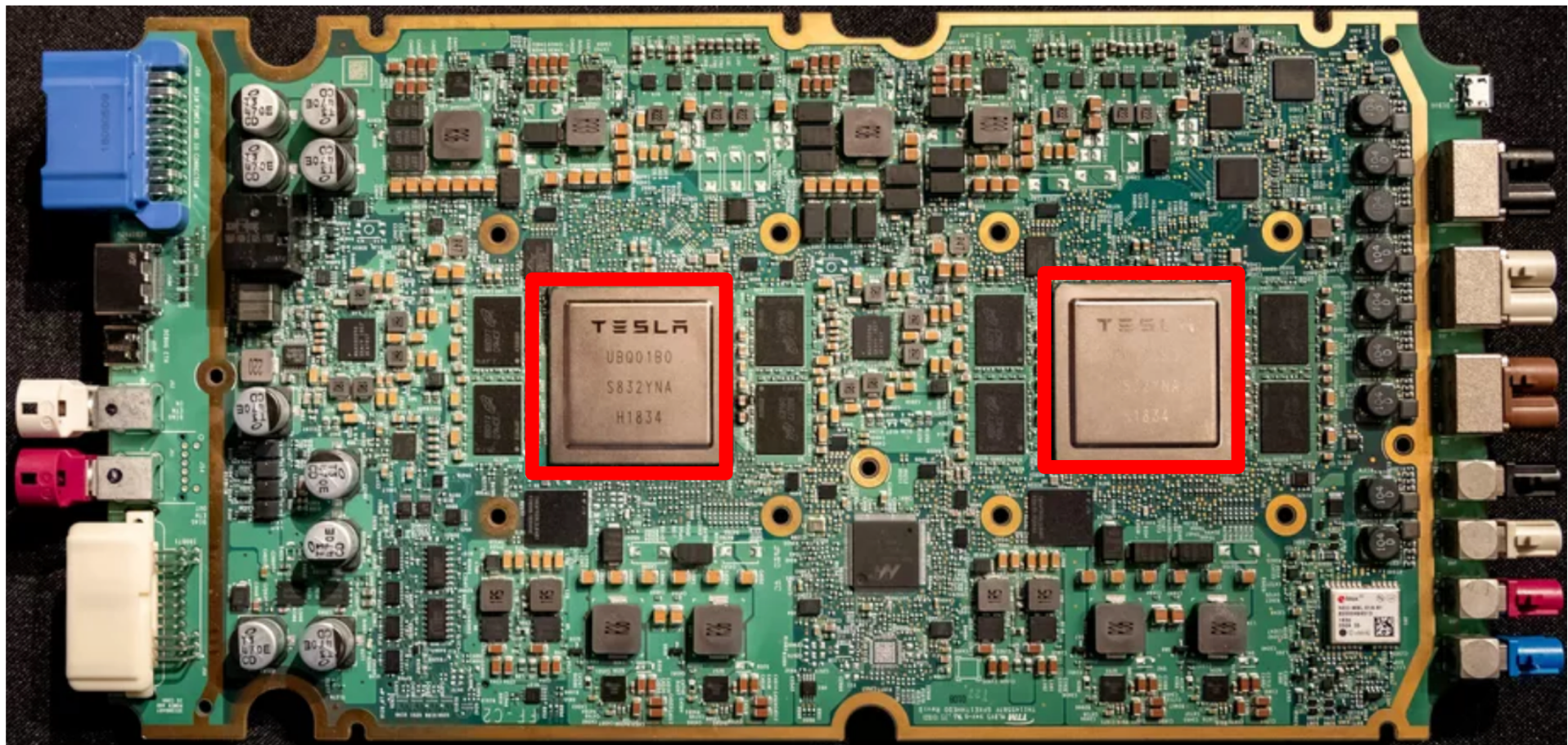


**Figure 4.** Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit”, ISCA 2017.

# Different Platforms, Different Goals

- ML accelerator: 260 mm<sup>2</sup>, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.

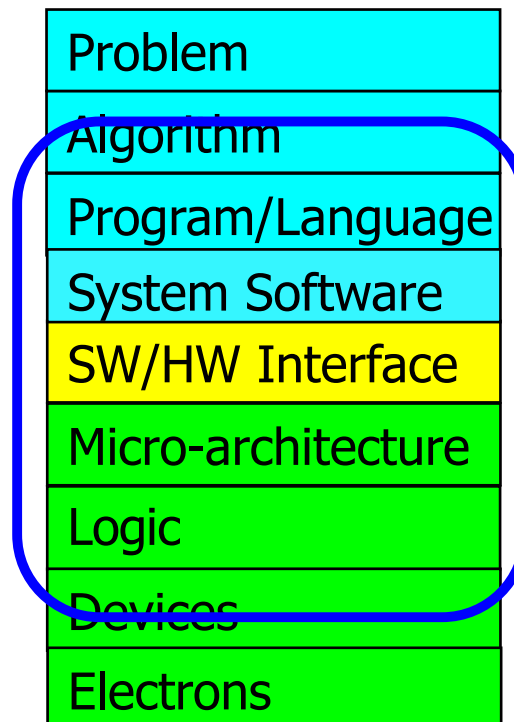


# Axiom

---

To achieve the highest **energy efficiency** and **performance**:

**we must take the expanded view**  
of computer architecture



**Co-design across the hierarchy:**  
**Algorithms to devices**

**Specialize as much as possible**  
**within the design goals**

Many Interesting Things  
Are Happening Today  
in Computer Architecture

Many Interesting Things  
Are Happening Today  
in Computer Architecture

**Performance  
and  
Energy Efficiency**

# Intel Optane Persistent Memory (2019)

---

- Non-volatile main memory
- Based on 3D-XPoint Technology



# PCM as Main Memory: Idea in 2009

---

- Benjamin C. Lee, Engin Ipek, Onur Mutlu, and Doug Burger, **"Architecting Phase Change Memory as a Scalable DRAM Alternative"**  
*Proceedings of the 36th International Symposium on Computer Architecture (ISCA)*, pages 2-13, Austin, TX, June 2009. [Slides \(pdf\)](#)

## Architecting Phase Change Memory as a Scalable DRAM Alternative

Benjamin C. Lee† Engin Ipek† Onur Mutlu‡ Doug Burger†

†Computer Architecture Group  
Microsoft Research  
Redmond, WA  
{blee, ipek, dburger}@microsoft.com

‡Computer Architecture Laboratory  
Carnegie Mellon University  
Pittsburgh, PA  
onur@cmu.edu

# PCM as Main Memory: Idea in 2009

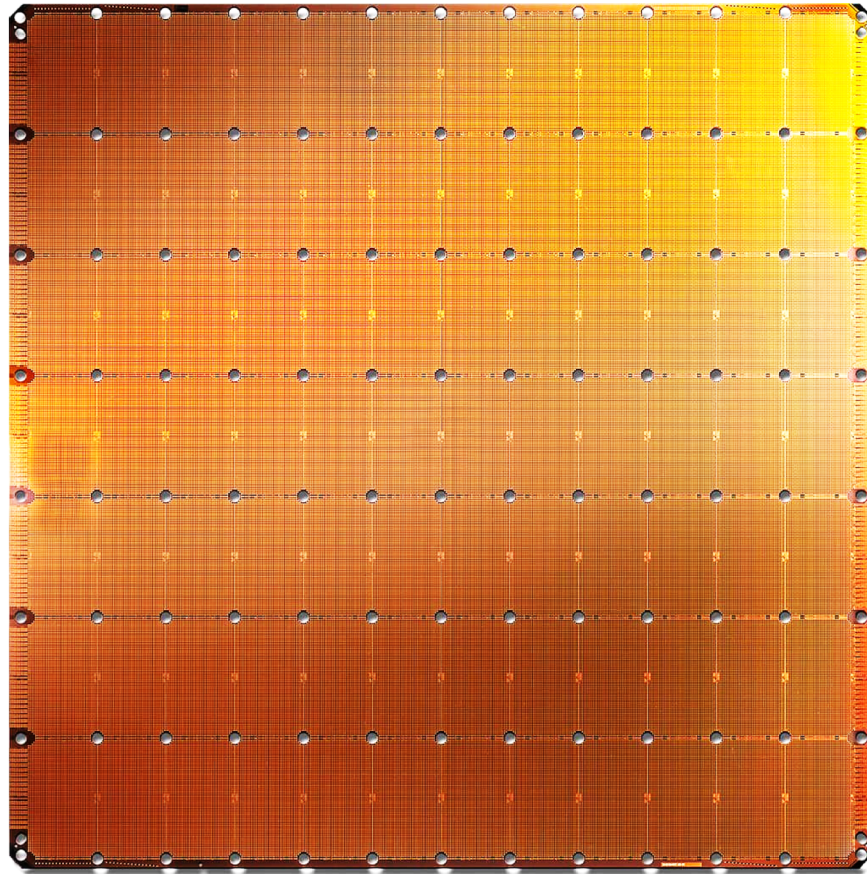
---

- Benjamin C. Lee, Ping Zhou, Jun Yang, Youtao Zhang, Bo Zhao, Engin Ipek, Onur Mutlu, and Doug Burger,  
**"Phase Change Technology and the Future of Main Memory"**  
*IEEE Micro*, Special Issue: Micro's Top Picks from 2009 Computer Architecture Conferences (**MICRO TOP PICKS**), Vol. 30, No. 1, pages 60-70, January/February 2010.

## PHASE-CHANGE TECHNOLOGY AND THE FUTURE OF MAIN MEMORY

# Cerebras's Wafer Scale Engine (2019)

---



## **Cerebras WSE**

1.2 Trillion transistors  
46,225 mm<sup>2</sup>

- The largest ML accelerator chip
- 400,000 cores



## **Largest GPU**

21.1 Billion transistors  
815 mm<sup>2</sup>

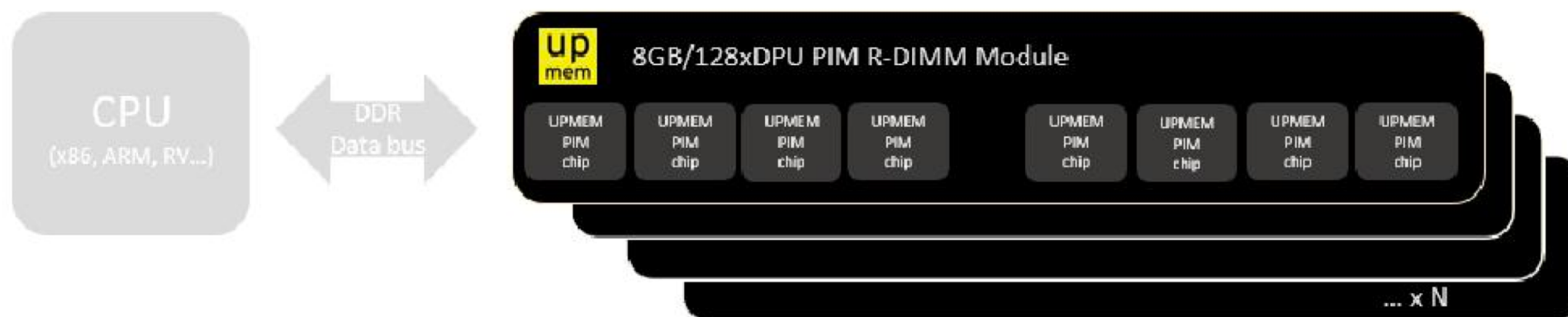
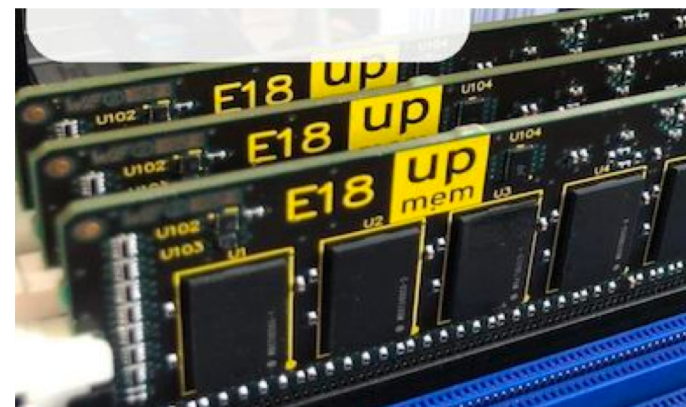
NVIDIA TITAN V

<https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning>

<https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning>

# UPMEM Processing-in-DRAM Engine (2019)

- **Processing in DRAM Engine**
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.
- Replaces **standard DIMMs**
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



# Specialized Processing in Memory (2015)

---

- Junwhan Ahn, Sungpack Hong, Sungjoo Yoo, Onur Mutlu, and Kiyoungh Choi,  
**"A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing"**  
*Proceedings of the 42nd International Symposium on Computer Architecture (ISCA)*, Portland, OR, June 2015.  
[[Slides \(pdf\)](#)] [[Lightning Session Slides \(pdf\)](#)]

## A Scalable Processing-in-Memory Accelerator for Parallel Graph Processing

Junwhan Ahn   Sungpack Hong<sup>§</sup>   Sungjoo Yoo   Onur Mutlu<sup>†</sup>   Kiyoungh Choi

junwhan@snu.ac.kr, sungpack.hong@oracle.com, sungjoo.yoo@gmail.com, onur@cmu.edu, kchoi@snu.ac.kr

Seoul National University

<sup>§</sup>Oracle Labs

<sup>†</sup>Carnegie Mellon University

# Processing in Memory on Mobile Devices

---

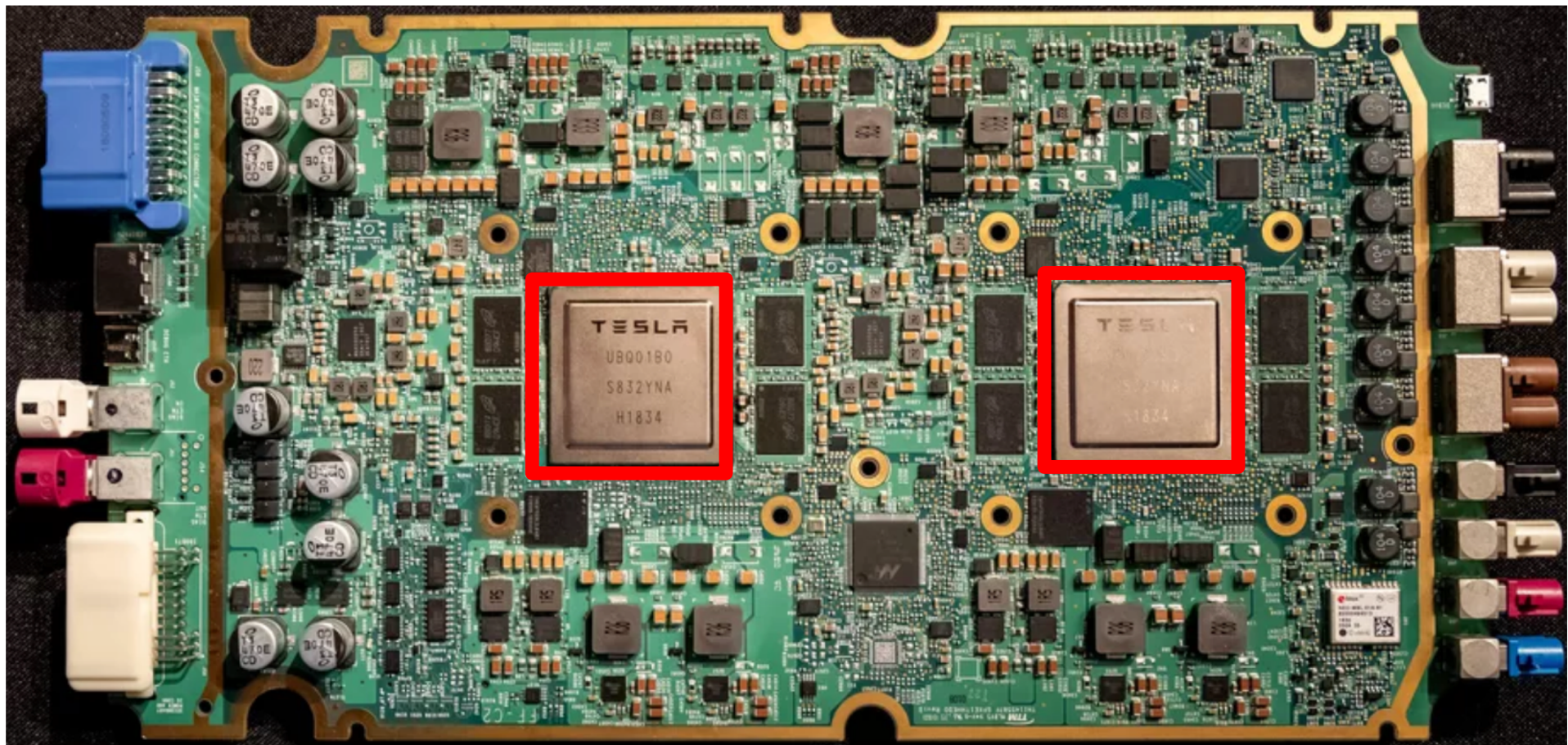
- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, **"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"**  
*Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Williamsburg, VA, USA, March 2018.

## Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

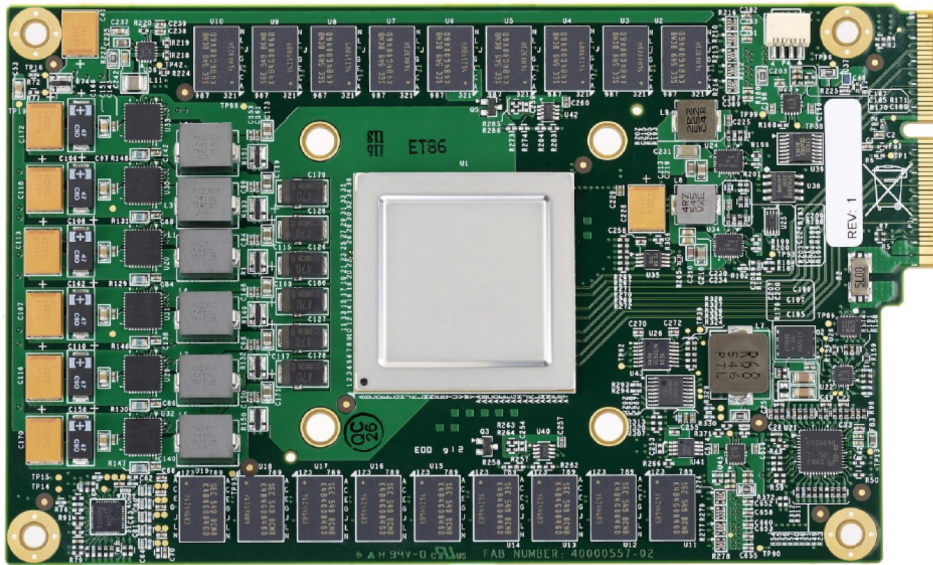
Amirali Boroumand <sup>1</sup>	Saugata Ghose <sup>1</sup>	Youngsok Kim <sup>2</sup>	
Rachata Ausavarungnirun <sup>1</sup>	Eric Shiu <sup>3</sup>	Rahul Thakur <sup>3</sup>	Daehyun Kim <sup>4,3</sup>
Aki Kuusela <sup>3</sup>	Allan Knies <sup>3</sup>	Parthasarathy Ranganathan <sup>3</sup>	Onur Mutlu <sup>5,1</sup>

# TESLA Full Self-Driving Computer (2019)

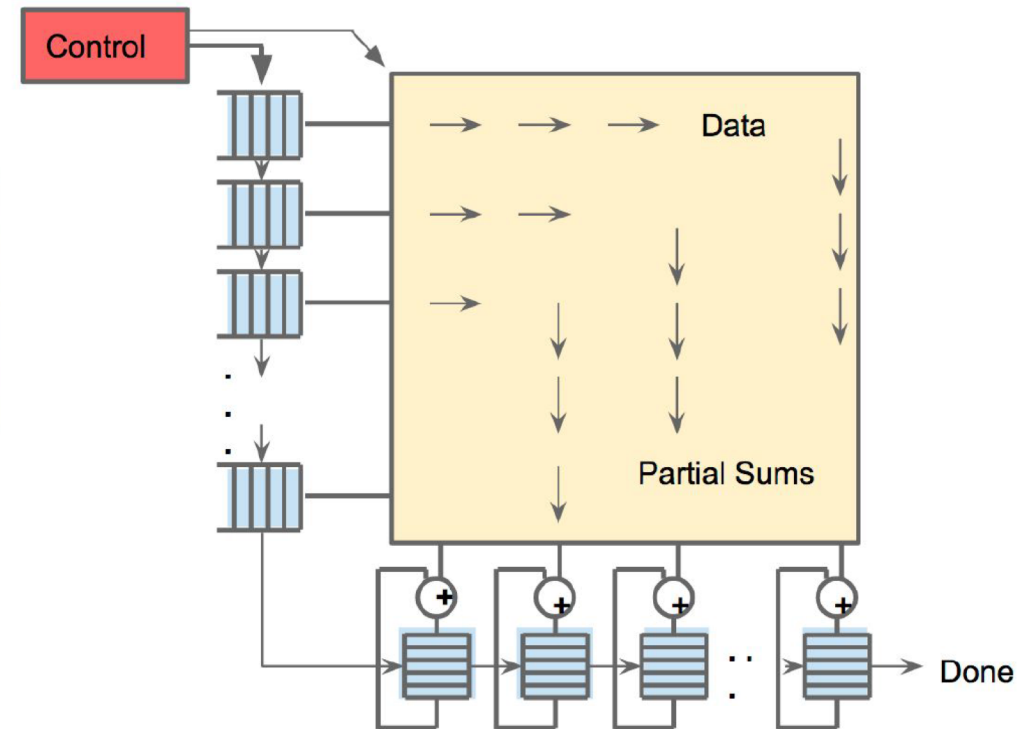
- ML accelerator: 260 mm<sup>2</sup>, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.



# Google TPU Generation I (~2016)



**Figure 3.** TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

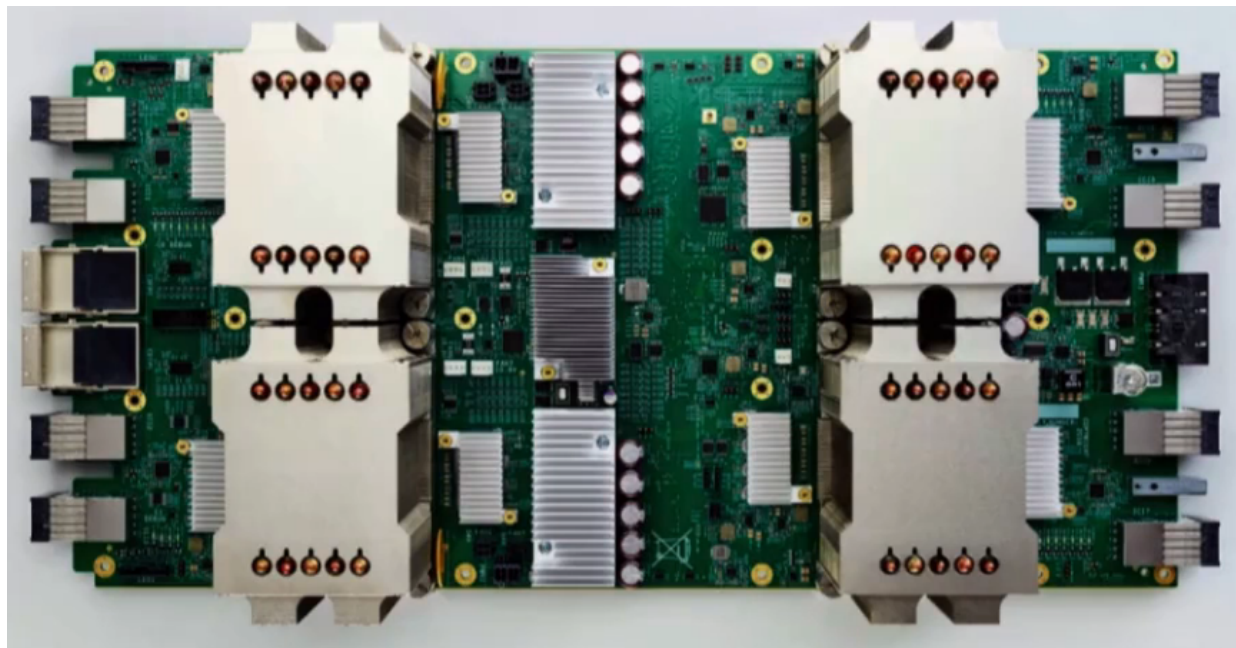


**Figure 4.** Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit”, ISCA 2017.

# Google TPU Generation II (2017)

---



<https://www.nextplatform.com/2017/05/17/first-depth-look-googles-new-second-generation-tpu/>

4 TPU chips

vs 1 chip in TPU1

High Bandwidth Memory

vs DDR3

Floating point operations

vs FP16

45 TFLOPS per chip

vs 23 TOPS

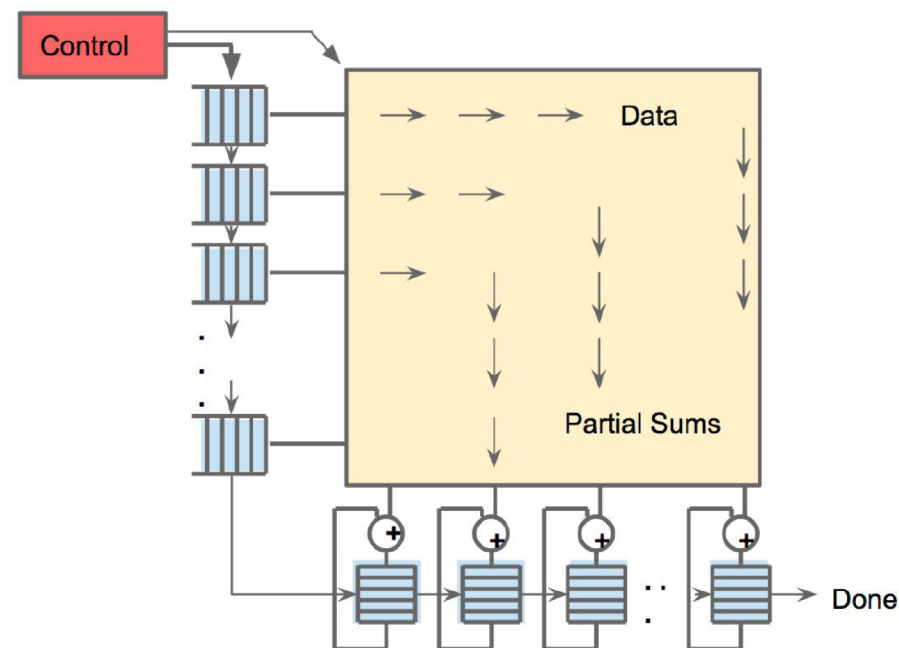
Designed for training

and inference

vs only inference

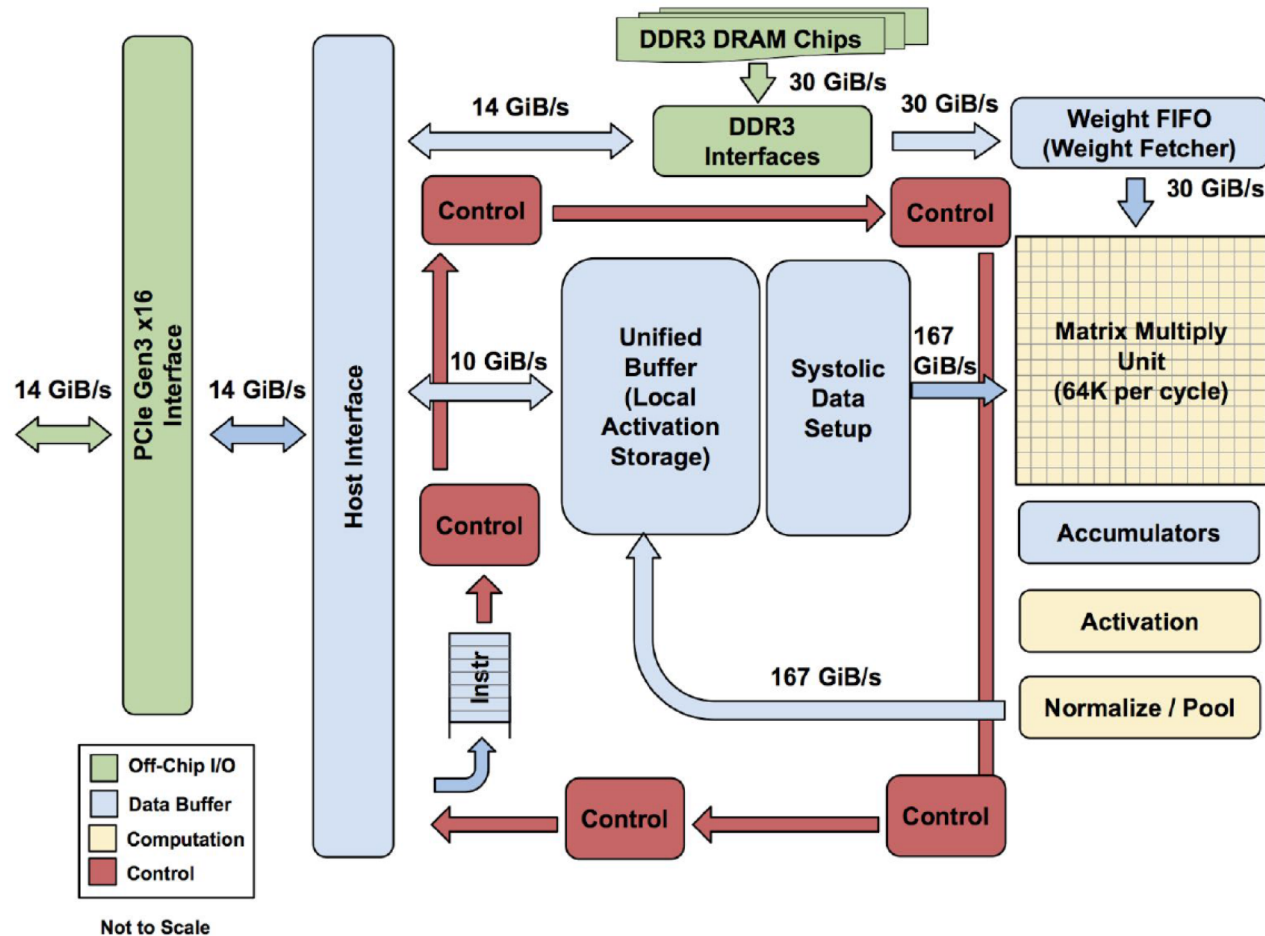
# An Example Modern Systolic Array: TPU (II)

As reading a large SRAM uses much more power than arithmetic, the matrix unit uses systolic execution to save energy by reducing reads and writes of the Unified Buffer [Kun80][Ram91][Ovt15b]. Figure 4 shows that data flows in from the left, and the weights are loaded from the top. A given 256-element multiply-accumulate operation moves through the matrix as a diagonal wavefront. The weights are preloaded, and take effect with the advancing wave alongside the first data of a new block. Control and data are pipelined to give the illusion that the 256 inputs are read at once, and that they instantly update one location of each of 256 accumulators. From a correctness perspective, software is unaware of the systolic nature of the matrix unit, but for performance, it does worry about the latency of the unit.



Jouppi et al., “[In-Datcenter Performance Analysis of a Tensor Processing Unit](#)”, ISCA 2017.

# An Example Modern Systolic Array: TPU (III)



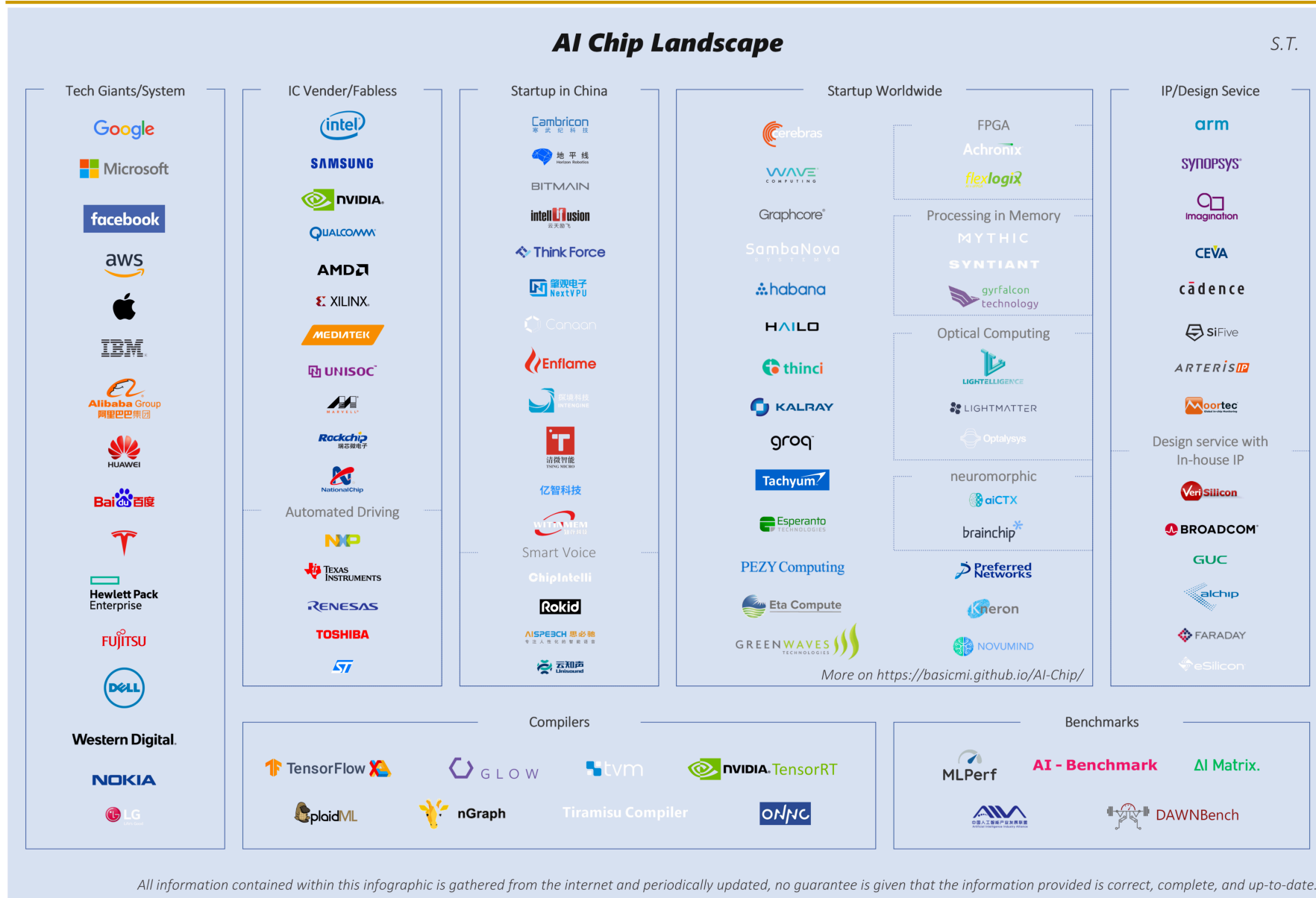
**Figure 1.** TPU Block Diagram. The main computation part is the yellow Matrix Multiply unit in the upper right hand corner. Its inputs are the blue Weight FIFO and the blue Unified Buffer (UB) and its output is the blue Accumulators (Acc). The yellow Activation Unit performs the nonlinear functions on the Acc, which go to the UB.

# Many (Other) AI/ML Chips

---

- Alibaba
- Amazon
- Facebook
- Google
- Huawei
- Intel
- Microsoft
- NVIDIA
- Tesla
- Many Others and Many Startups...
- **Many More to Come...**

# Many (Other) AI/ML Chips



Many Interesting Things  
Are Happening Today  
in Computer Architecture

Many Interesting Things  
Are Happening Today  
in Computer Architecture

**Reliability  
and  
Security**

# Security: RowHammer (2014)

---



# The Story of RowHammer

- One can **predictably induce bit flips** in commodity DRAM chips
  - >80% of the tested DRAM chips are vulnerable
- First example of how a **simple hardware failure mechanism** can create a **widespread system security vulnerability**

**WIRED**

Forget Software—Now Hackers Are Exploiting Physics

BUSINESS

CULTURE

DESIGN

GEAR

SCIENCE

ANDY GREENBERG SECURITY 08.31.16 7:00 AM

SHARE



SHARE  
18276



TWEET

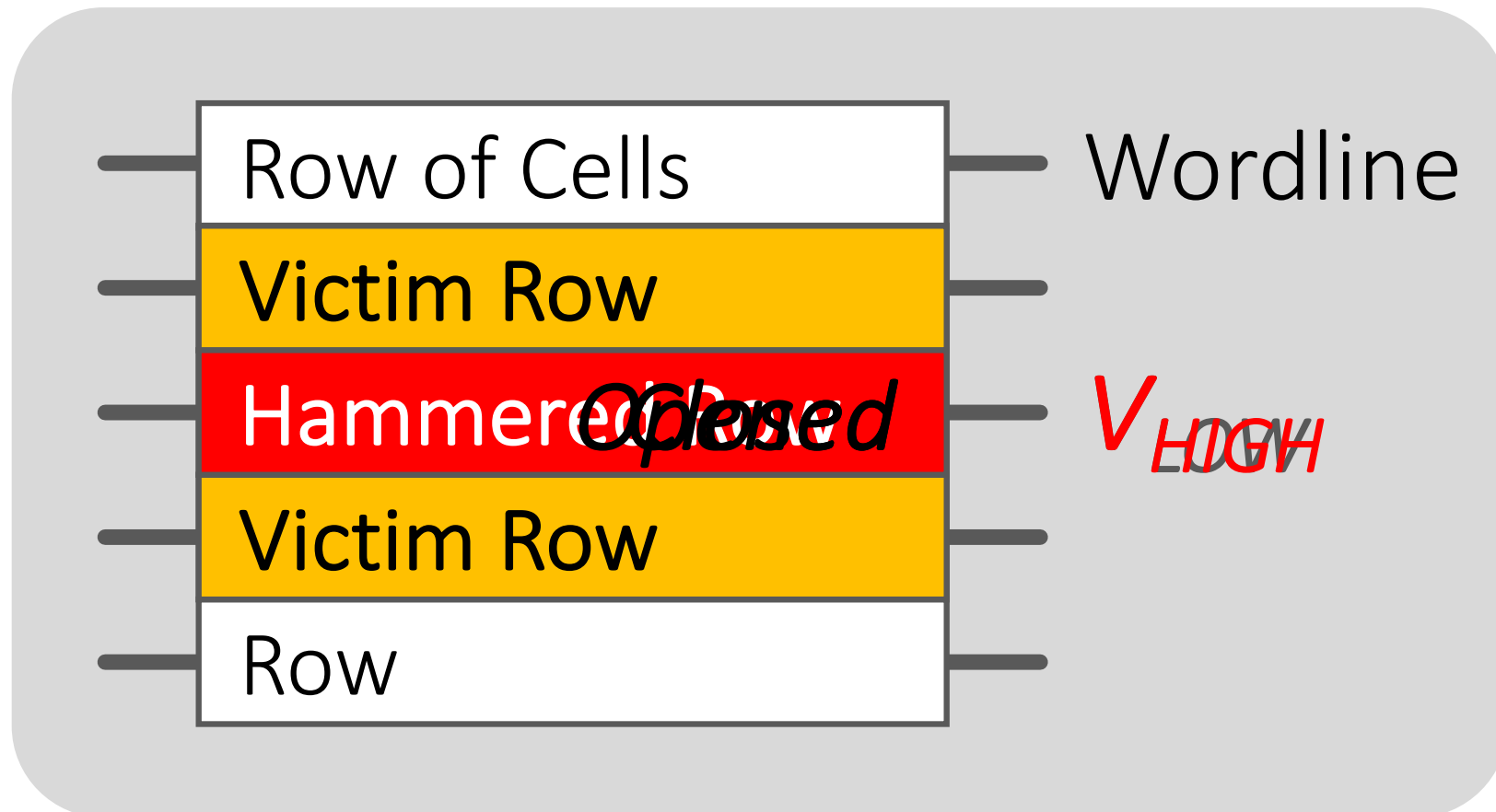
# FORGET SOFTWARE—NOW HACKERS ARE EXPLOITING PHYSICS

# Security: RowHammer (2014)



It's like breaking into an apartment by repeatedly slamming a neighbor's door until the vibrations open the door you were after

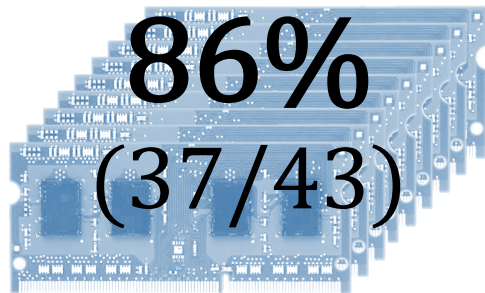
# Modern DRAM is Prone to Disturbance Errors



Repeatedly reading a row enough times (before memory gets refreshed) induces **disturbance errors** in adjacent rows in **most real DRAM chips you can buy today**

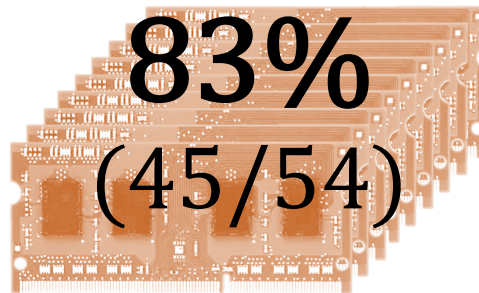
# Most DRAM Modules Are Vulnerable

A company



Up to  
 $1.0 \times 10^7$   
errors

B company



Up to  
 $2.7 \times 10^6$   
errors

C company



Up to  
 $3.3 \times 10^5$   
errors

# RowHammer: Five Years Ago...

---

- Yoongu Kim, Ross Daly, Jeremie Kim, Chris Fallin, Ji Hye Lee, Donghyuk Lee, Chris Wilkerson, Konrad Lai, and Onur Mutlu,  
**"Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors"**  
*Proceedings of the 41st International Symposium on Computer Architecture (ISCA)*, Minneapolis, MN, June 2014.  
[\[Slides \(pptx\) \(pdf\)\]](#) [\[Lightning Session Slides \(pptx\) \(pdf\)\]](#) [\[Source Code and Data\]](#)

## Flipping Bits in Memory Without Accessing Them: An Experimental Study of DRAM Disturbance Errors

Yoongu Kim<sup>1</sup>   Ross Daly\*   Jeremie Kim<sup>1</sup>   Chris Fallin\*   Ji Hye Lee<sup>1</sup>  
Donghyuk Lee<sup>1</sup>   Chris Wilkerson<sup>2</sup>   Konrad Lai   Onur Mutlu<sup>1</sup>

<sup>1</sup>Carnegie Mellon University   <sup>2</sup>Intel Labs

# RowHammer: Now and Beyond...

---

- Onur Mutlu and Jeremie Kim,  
**"RowHammer: A Retrospective"**  
*IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD) Special Issue on Top Picks in Hardware and Embedded Security*, 2019.  
[[Preliminary arXiv version](#)]

## RowHammer: A Retrospective

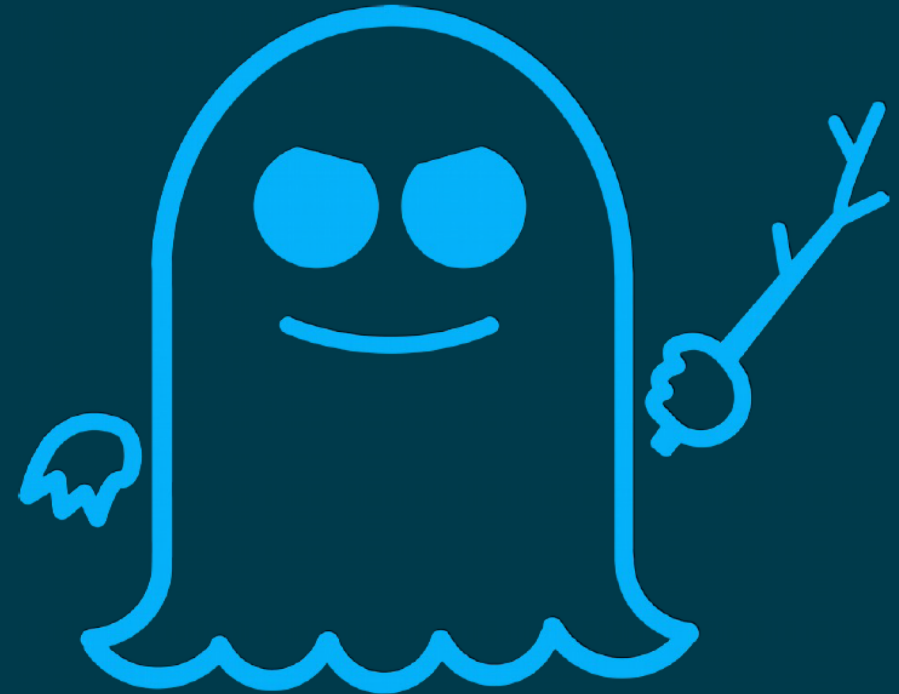
Onur Mutlu<sup>§‡</sup>      Jeremie S. Kim<sup>‡§</sup>  
§ETH Zürich      ‡Carnegie Mellon University

# Security: Meltdown and Spectre (2018)

---



**MELTDOWN**



**SPECTRE**

# Meltdown and Spectre

---

- Someone can steal secret data from the system even though
  - your program and data are perfectly correct and
  - your hardware behaves according to the specification and
  - there are no software vulnerabilities/bugs
  
- Why?
  - Speculative execution leaves traces of secret data in the processor's cache (internal storage)
    - It brings data that is not supposed to be brought/accessed if there was no speculative execution
  - A malicious program can inspect the contents of the cache to "infer" secret data that it is not supposed to access
  - A malicious program can actually force another program to speculatively execute code that leaves traces of secret data

# More on Meltdown/Spectre Vulnerabilities

---

## Project Zero

News and updates from the Project Zero team at Google

Wednesday, January 3, 2018

### Reading privileged memory with a side-channel

Posted by Jann Horn, Project Zero

We have discovered that CPU data cache timing can be abused to efficiently leak information out of mis-speculated execution, leading to (at worst) arbitrary virtual memory read vulnerabilities across local security boundaries in various contexts.

Many Interesting Things  
Are Happening Today  
in Computer Architecture

Many Interesting Things  
Are Happening Today  
in Computer Architecture

**More Demanding Workloads**

# New Genome Sequencing Technologies

---

## Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

*Briefings in Bioinformatics*, bby017, <https://doi.org/10.1093/bib/bby017>

**Published:** 02 April 2018    **Article history** ▼



Oxford Nanopore MinION

Data → performance & energy bottleneck

# Why Do We Care? An Example

200 Oxford Nanopore sequencers have left UK for China, to support rapid, near-sample coronavirus sequencing for outbreak surveillance

Fri 31st January 2020

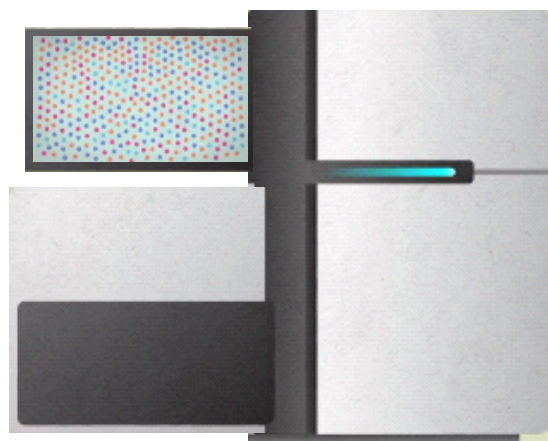
Following extensive support of, and collaboration with, public health professionals in China, Oxford Nanopore has shipped an additional 200 MinION sequencers and related consumables to China. These will be used to support the ongoing surveillance of the current coronavirus outbreak, adding to a large number of the devices already installed in the country.



Each MinION sequencer is approximately the size of a stapler, and can provide rapid sequence information about the coronavirus.

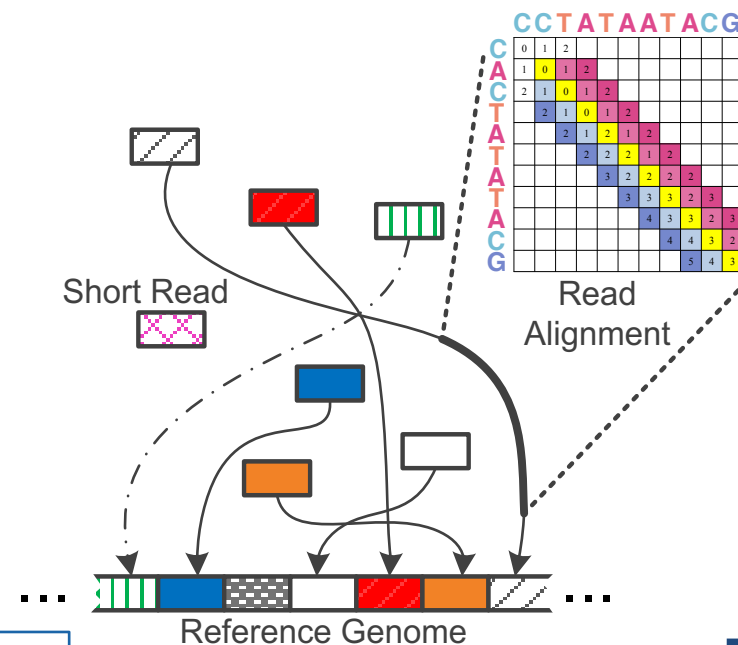


700Kg of Oxford Nanopore sequencers and consumables are on their way for use by Chinese scientists in understanding the current coronavirus outbreak.



Billions of Short Reads

TATATATACGTACTAGTACGT  
 TTTAGTACGTACGT  
 ATACGTACTAGTACGT  
 ACGCCCCTACGTA  
 ACGTACTAGTACGT  
 TTAGTACGTACGT  
 TACGTACTAAAGTACGT  
 TACGTACTAGTACGT  
 TTTAAACGTA  
 CGTACTAGTACGT  
 GGGAGTACGTACGT



## 1 Sequencing

# Genome Analysis

## 2 Read Mapping

Data → performance & energy bottleneck

read4: CGCTTCCAT  
 read5: CCATGACGC  
 read6: TTCCATGAC

## 3 Variant Calling

## 4 Scientific Discovery

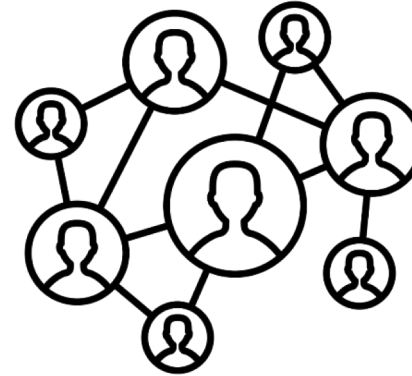


# Data Overwhelms Modern Machines

---



**In-memory Databases**



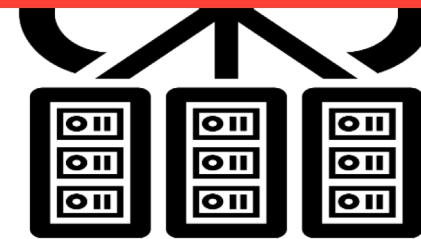
**Graph/Tree Processing**

**Data → performance & energy bottleneck**



**In-Memory Data Analytics**

[Clapp+ (Intel), IISWC'15;  
Awan+, BDCloud'15]



**Datacenter Workloads**

[Kanev+ (Google), ISCA'15]

# Data Overwhelms Modern Machines



**Chrome**



**TensorFlow Mobile**

Data → performance & energy bottleneck

**VP9**



**Video Playback**

Google's **video codec**

**VP9**



**Video Capture**

Google's **video codec**

# Data Movement Overwhelms Modern Machines

- Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, **"Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks"** *Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Williamsburg, VA, USA, March 2018.

**62.7% of the total system energy  
is spent on data movement**

## Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks

Amirali Boroumand<sup>1</sup>      Saugata Ghose<sup>1</sup>      Youngsok Kim<sup>2</sup>  
Rachata Ausavarungnirun<sup>1</sup>      Eric Shiu<sup>3</sup>      Rahul Thakur<sup>3</sup>      Daehyun Kim<sup>4,3</sup>  
Aki Kuusela<sup>3</sup>      Allan Knies<sup>3</sup>      Parthasarathy Ranganathan<sup>3</sup>      Onur Mutlu<sup>5,1</sup>

Many Interesting Things  
Are Happening Today  
in Computer Architecture

# Many Novel Concepts Investigated Today

---

## ■ New Computing Paradigms (Rethinking the Full Stack)

- ❑ Processing in Memory, Processing Near Data
- ❑ Neuromorphic Computing
- ❑ Fundamentally Secure and Dependable Computers

## ■ New Accelerators (Algorithm-Hardware Co-Designs)

- ❑ Artificial Intelligence & Machine Learning
- ❑ Graph Analytics
- ❑ Genome Analysis

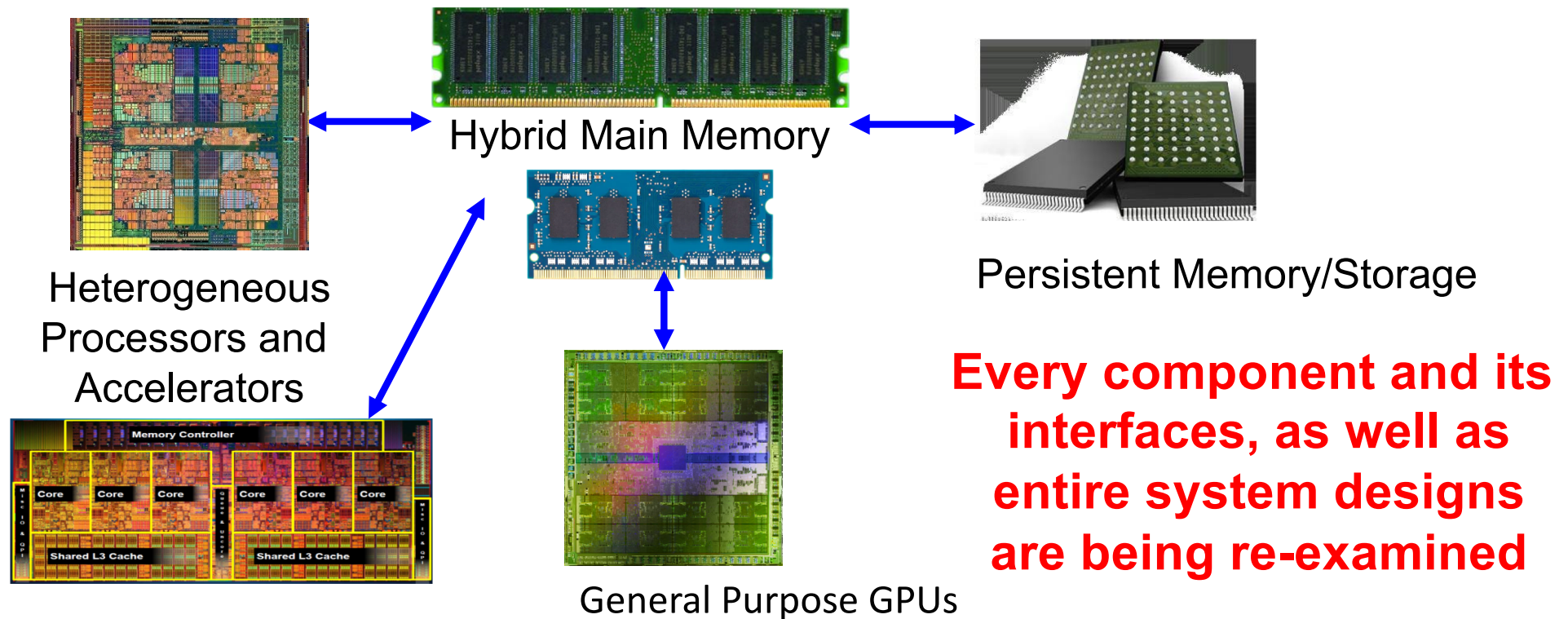
## ■ New Memories and Storage Systems

- ❑ Non-Volatile Main Memory
- ❑ Intelligent Memory

# Computer Architecture Today

---

- Computing landscape is very different from 10-20 years ago
- Applications and technology both demand novel architectures



# Computer Architecture Today (II)

---

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)
- You can invent new paradigms for computation, communication, and storage
- Recommended book: Thomas Kuhn, “[The Structure of Scientific Revolutions](#)” (1962)
  - Pre-paradigm science: no clear consensus in the field
  - Normal science: dominant theory used to explain/improve things (business as usual); exceptions considered anomalies
  - Revolutionary science: underlying assumptions re-examined

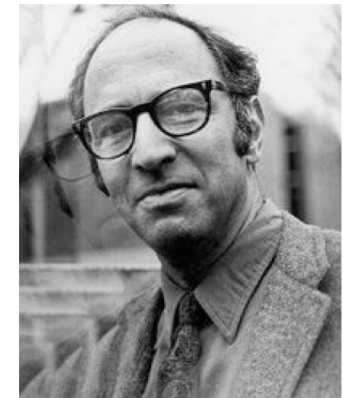
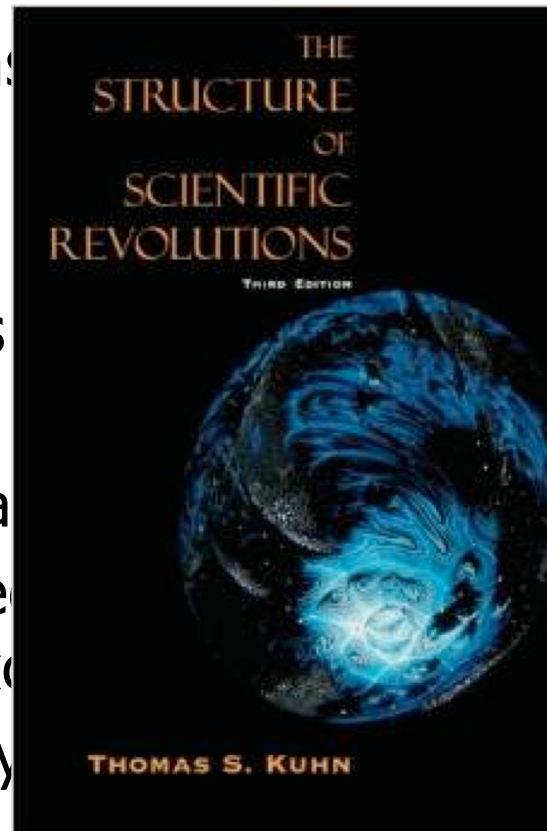
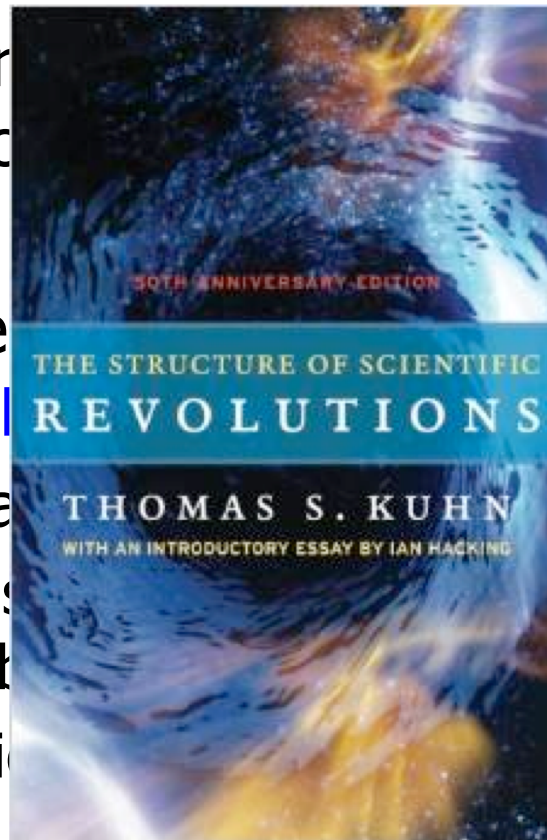
# Computer Architecture Today (II)

- You can revolutionize the way computers are built, if you understand both the hardware and the software (and change each accordingly)

- You can improve communication

- Recommended reading: **Scientific Revolutions** (1962)

- Pre-para
- Normal s
- things (b
- Revolution



ure of

field  
improve  
anomalies  
examined

# Takeaways

---

- It is an exciting time to be understanding and designing computing architectures
- Many challenging and exciting problems in platform design
  - That no one has tackled (or thought about) before
  - That can have huge impact on the world's future
- Driven by huge hunger for data (Big Data), new applications (ML/AI, graph analytics, genomics), ever-greater realism, ...
  - We can easily collect more data than we can analyze/understand
- Driven by significant difficulties in keeping up with that hunger at the technology layer
  - Five walls: Energy, reliability, complexity, security, scalability

# Let's Start with Some Puzzles

a.k.a. Computer Architecture resembles Building Architecture

# What Is This?



# What About This?



What Do the Following  
Have in Common?

# Gare do Oriente, Lisbon

---



# Milwaukee Art Museum



# Athens Olympic Stadium

---



# City of Arts and Sciences, Valencia

---



# Florida Polytechnic University (I)

---



# Oculus, New York City

---



Source: <https://www.dezeen.com/2016/08/29/santiago-calatrava-oculus-world-trade-center-transportation-hub-new-york-photographs-hufton-crow/>

What do All Those Have in Common  
with Bahnhof Stadelhofen?

# Answer: All Designed by a Famous Architect

---

- ETH Alumnus, PhD Civil Engineering
- “The train station has several of the features that became signatures of his work; straight lines and right angles are rare.”



**Santiago Calatrava Valls** (born 28 July 1951) is a Spanish [architect](#), [structural engineer](#), [sculptor](#) and [painter](#), particularly known for his bridges supported by single leaning pylons, and his railway stations, stadiums, and museums, whose sculptural forms often resemble living organisms.<sup>[1]</sup> His best-known works include the [Milwaukee Art Museum](#), the [Turning Torso](#) tower in [Malmo](#), Sweden, the [Margaret Hunt Hill Bridge](#) in [Dallas, Texas](#), and the [Museum of Tomorrow](#) in [Rio de Janeiro](#),

# Your First Comp. Architecture Assignment

---

- Go and find the closest Calatrava building to this classroom
  - For the ones who like a challenge, find the furthest building that was designed by Calatrava to his classroom 😊
- Appreciate the beauty & out-of-the-box and creative thinking
- Think about tradeoffs in the design
  - Strengths, weaknesses, goals of design
- Derive principles on your own for good design and innovation
- Due date: **Any time during this course**
  - Later during the course is better
  - Apply what you have learned in this course
  - Think out-of-the-box

# But First, Today's First Assignment

---

# Find The Differences of This and That

# This



# That

---



# Many Tradeoffs Between Two Designs

---

- You can list them after you complete the first assignment...

# Aside: Evaluation Criteria for the Designs

---

- Functionality (Does it meet the specification?)
  - Reliability
  - Space requirement
  - Cost
  - Expandability
  - Comfort level of users
  - Happiness level of users
  - Aesthetics
  - Security
  - ...
- 
- How to evaluate goodness of design is always a critical question → "Performance" evaluation and metrics

# A Key Question

---

- How was Calavatra able to design especially his key buildings?
- Can have many guesses
  - (Ultra) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, ...)
  - Funding (\$\$\$\$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - ...
- (You will be exposed to and hopefully develop/enhance many of these skills in this course)

# Principled Design

---

- “To me, there are **two overriding principles** to be found in nature which are most appropriate for building:
  - one is the **optimal use of material**,
  - the other **the capacity of organisms to change shape, to grow, and to move.**”
  - *Santiago Calatrava*
  
- “Calatrava's constructions are inspired by natural forms like plants, bird wings, and the human body.”

# Gare do Oriente, Lisbon, Revisited



Source: By Martín Gómez Tagle - Lisbon, Portugal, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=13764903>

Source: <http://www.arcspace.com/exhibitions/unsorted/santiago-calatrava/>

# A Principled Design

---

## Zoomorphic architecture

---

From Wikipedia, the free encyclopedia

**Zoomorphic architecture** is the practice of using animal forms as the inspirational basis and blueprint for architectural design. "While animal forms have always played a role adding some of the deepest layers of meaning in architecture, it is now becoming evident that a new strand of **biomorphism** is emerging where the meaning derives not from any specific representation but from a more general allusion to biological processes."<sup>[1]</sup>

Some well-known examples of Zoomorphic architecture can be found in the **TWA Flight Center** building in **New York City**, by **Eero Saarinen**, or the **Milwaukee Art Museum** by **Santiago Calatrava**, both inspired by the form of a bird's wings.<sup>[3]</sup>

# What Does This Remind You Of?

---



# The Architect's Answer

---

## Design [\[ edit \]](#)

Calatrava said that the Oculus resembles a bird being released from a child's hand. The roof was originally designed to mechanically open to increase light and ventilation to the enclosed space. [Herbert Muschamp](#), architecture critic of *The New York Times*, compared the design to the [Bethesda Terrace and Fountain](#) in [Central Park](#), and wrote in 2004:

# Strengths and Praise

---

“ Santiago Calatrava's design for the World Trade Center PATH station should satisfy those who believe that buildings planned for ground zero must aspire to a spiritual dimension. Over the years, many people have discerned a metaphysical element in Mr. Calatrava's work. I hope New Yorkers will detect its presence, too. With deep appreciation, I congratulate the Port Authority for commissioning Mr. Calatrava, the great Spanish architect and engineer, to design a building with the power to shape the future of New York. It is a pleasure to report, for once, that public officials are not overstating the case when they describe a design as breathtaking.<sup>[43]</sup>

”

# Design Constraints and Criticism

---

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

“ In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender *stegosaurus* more than it does a bird.<sup>[45]</sup> ”

# Stegosaurus

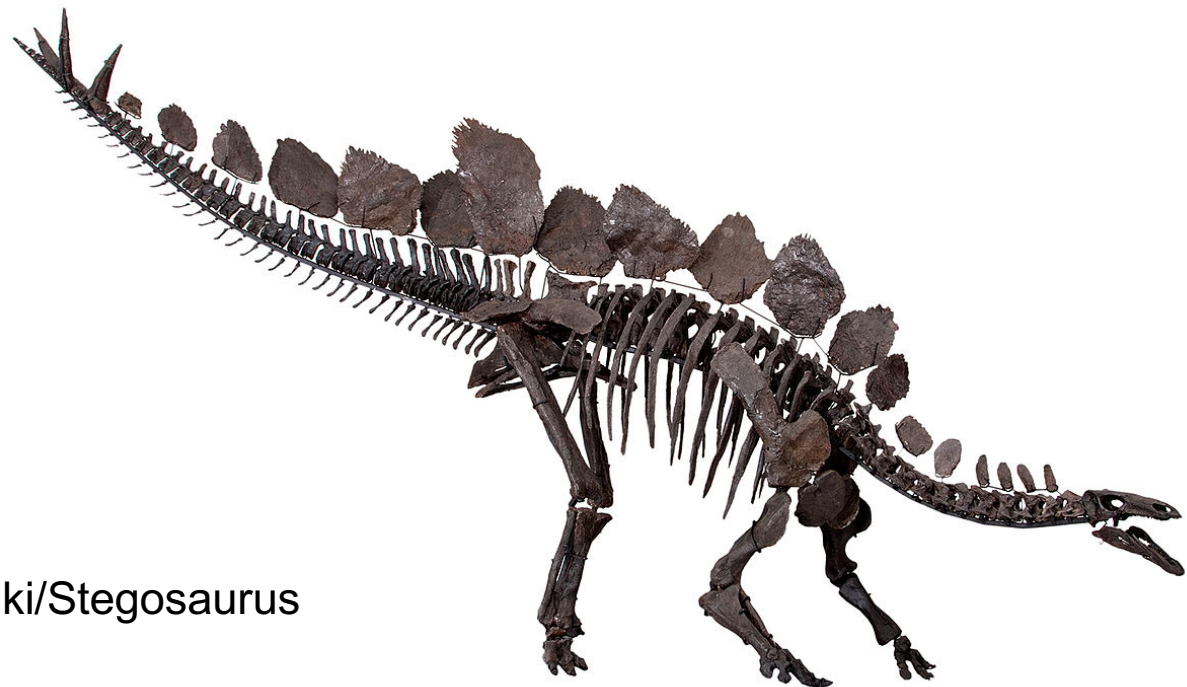
---

From Wikipedia, the free encyclopedia

---

For the *pachycephalosaurid* of a similar name, see *Stegoceras*.

**Stegosaurus** (/ˈstɛɡəˈsɔːrəs/<sup>[1]</sup>) is a genus of armored dinosaur. Fossils of this genus date to the Late Jurassic period, where they are found in Kimmeridgian to early Tithonian aged strata, between 155 and 150 million years ago, in the western United States and Portugal. Several



Source: <https://en.wikipedia.org/wiki/Stegosaurus>

Susannah Maidment et al. & Natural History Museum, London - Maidment SCR, Brassey C, Barrett PM (2015) The Postcranial Skeleton of an Exceptionally Complete Individual of the Plated Dinosaur *Stegosaurus stenops* (Dinosauria: Thyreophora) from the Upper Jurassic Morrison Formation of Wyoming, U.S.A. PLoS ONE 10(10): e0138352. doi:10.1371/journal.pone.0138352

# Design Constraints: Noone is Immune

---

However, Calatrava's original soaring spike design was scaled back because of security issues. The *New York Times* observed in 2005:

“ In the name of security, Santiago Calatrava's bird has grown a beak. Its ribs have doubled in number and its wings have lost their interstices of glass.... [T]he main transit hall, between Church and Greenwich Streets, will almost certainly lose some of its delicate quality, while gaining structural expressiveness. It may now evoke a slender *stegosaurus* more than it does a bird.<sup>[45]</sup> ”

The design was further modified in 2008 to eliminate the opening and closing roof mechanism because of budget and space constraints.<sup>[46]</sup>

The Transportation Hub has been dubbed "the world's most expensive transportation hub" for its massive cost for reconstruction—\$3.74 billion dollars.<sup>[48][58]</sup> By contrast, the proposed two-mile PATH extension

# The Lecture Was Slightly Different When I Was at CMU

# What Is This?



# Answer: Masterpiece of A Famous Architect

---

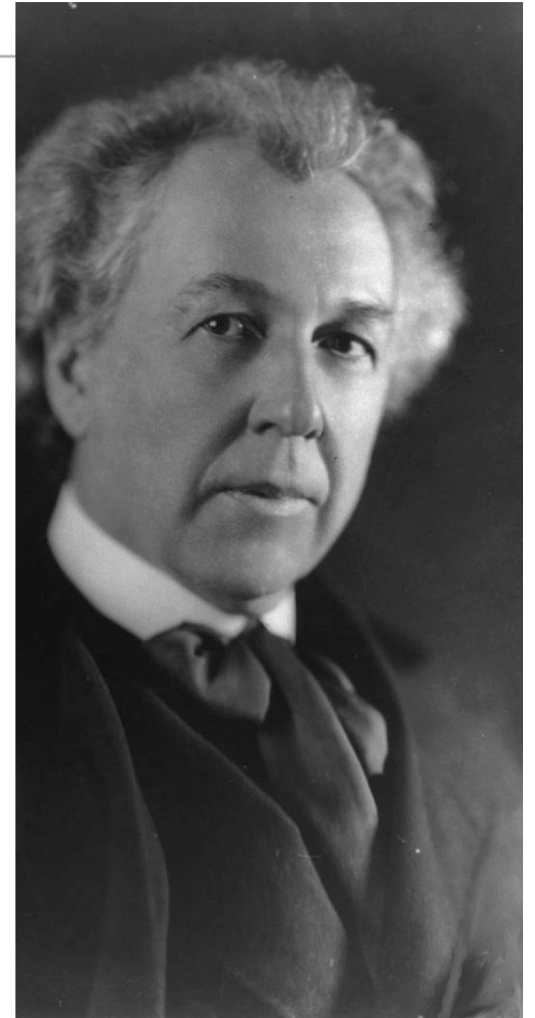
## Fallingwater

---

From Wikipedia, the free encyclopedia

**Fallingwater** or **Kaufmann Residence** is a house designed by architect [Frank Lloyd Wright](#) in 1935 in rural [southwestern Pennsylvania](#), 43 miles (69 km) southeast of [Pittsburgh](#).<sup>[4]</sup> The home was built partly over a waterfall on [Bear Run](#) in the Mill Run section of [Stewart Township, Fayette County, Pennsylvania](#), in the [Laurel Highlands](#) of the [Allegheny Mountains](#).

*Time* cited it after its completion as Wright's "most beautiful job";<sup>[5]</sup> it is listed among *Smithsonian's* Life List of 28 places "to visit before you die."<sup>[6]</sup> It was designated a [National Historic Landmark](#) in 1966.<sup>[3]</sup> In 1991, members of the [American Institute of Architects](#) named the house the "best all-time work of American architecture" and in 2007, it was ranked twenty-ninth on the [list of America's Favorite Architecture](#) according to the AIA.



# Find The Differences of This and That

# This

---



# That

---



# A Key Question

---

- How was Wright able to design his masterpiece?
- Can have many guesses
  - (Ultra) hard work, perseverance, dedication (over decades)
  - Experience
  - Creativity, Out-of-the-box thinking
  - A good understanding of past designs
  - Good judgment and intuition
  - Strong skill combination (math, architecture, art, engineering, ...)
  - Funding (\$\$\$\$), luck, initiative, entrepreneurialism
  - Strong understanding of and commitment to fundamentals
  - Principled design
  - ...
- (You will be exposed to and hopefully develop/enhance many of these skills in this course)

# A Quote from The Architect Himself

---

- “architecture [...] based upon **principle**, and not upon **precedent**”



# A Principled Design

---

## Organic architecture

---

From Wikipedia, the free encyclopedia

**Organic architecture** is a [philosophy](#) of [architecture](#) which promotes harmony between human habitation and the natural world through design approaches so sympathetic and well integrated with its site, that buildings, furnishings, and surroundings become part of a unified, interrelated composition.

A well-known example of organic architecture is [Fallingwater](#), the residence Frank Lloyd Wright designed for the Kaufmann family in rural Pennsylvania. Wright had many choices to locate a home on this large site, but chose to place the home directly over the waterfall and creek creating a close, yet noisy dialog with the rushing water and the steep site. The horizontal striations of stone masonry with daring [cantilevers](#) of colored beige concrete blend with native rock outcroppings and the wooded environment.

# A Key Question

---

- How was Wright able to design his masterpiece?
- Can have many guesses
  - ❑ (Ultra) hard work, perseverance, dedication (over decades)
  - ❑ Experience
  - ❑ Creativity, Out-of-the-box thinking
  - ❑ A good understanding of past designs
  - ❑ Good judgment and intuition
  - ❑ Strong skill combination (math, architecture, art, engineering, ...)
  - ❑ Funding (\$\$\$\$), luck, initiative, entrepreneurialism
  - ❑ Strong understanding of and commitment to fundamentals
  - ❑ Principled design
  - ❑ ...
- (You will be exposed to and hopefully develop/enhance many of these skills in this course)

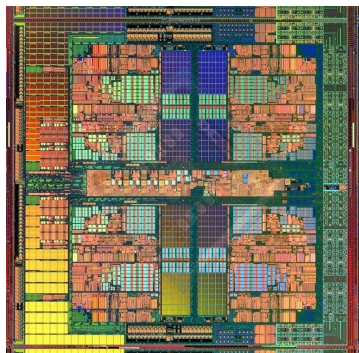
# Takeaways

---

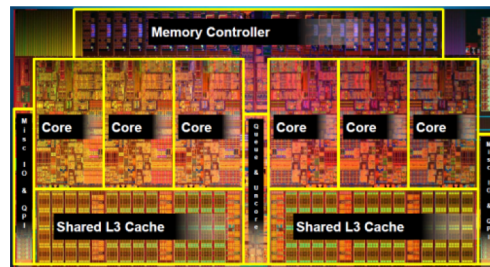
- It all starts from the basic building blocks and design principles
- And, knowledge of how to use & apply them
- Underlying technology might change (e.g., steel vs. wood)
  - but methods of taking advantage of technology bear resemblance
  - methods used for design depend on the principles employed

# The Same Applies to Processor Chips

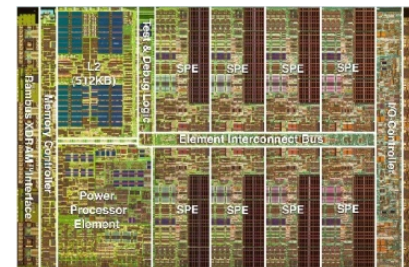
- There are **basic building blocks** and **design principles**



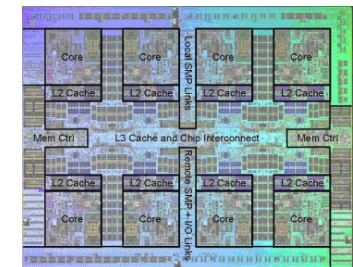
AMD Barcelona  
4 cores



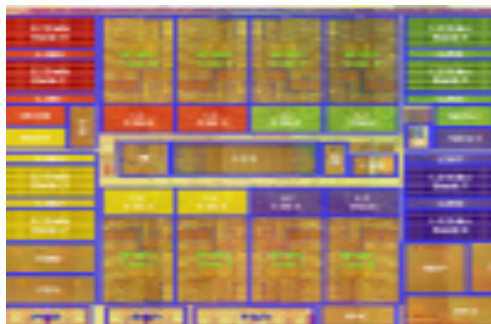
Intel Core i7  
8 cores



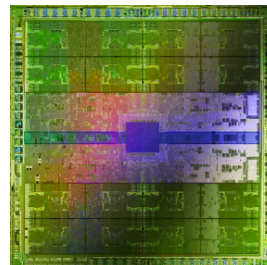
IBM Cell BE  
8+1 cores



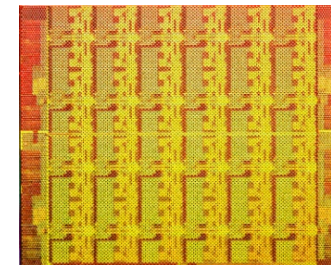
IBM POWER7  
8 cores



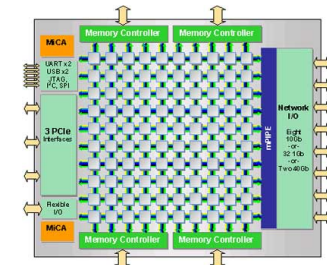
Sun Niagara II  
8 cores



Nvidia Fermi  
448 "cores"



Intel SCC  
48 cores, networked



Tiler TILE Gx  
100 cores, networked

# The Same Applies to Computing Systems

---

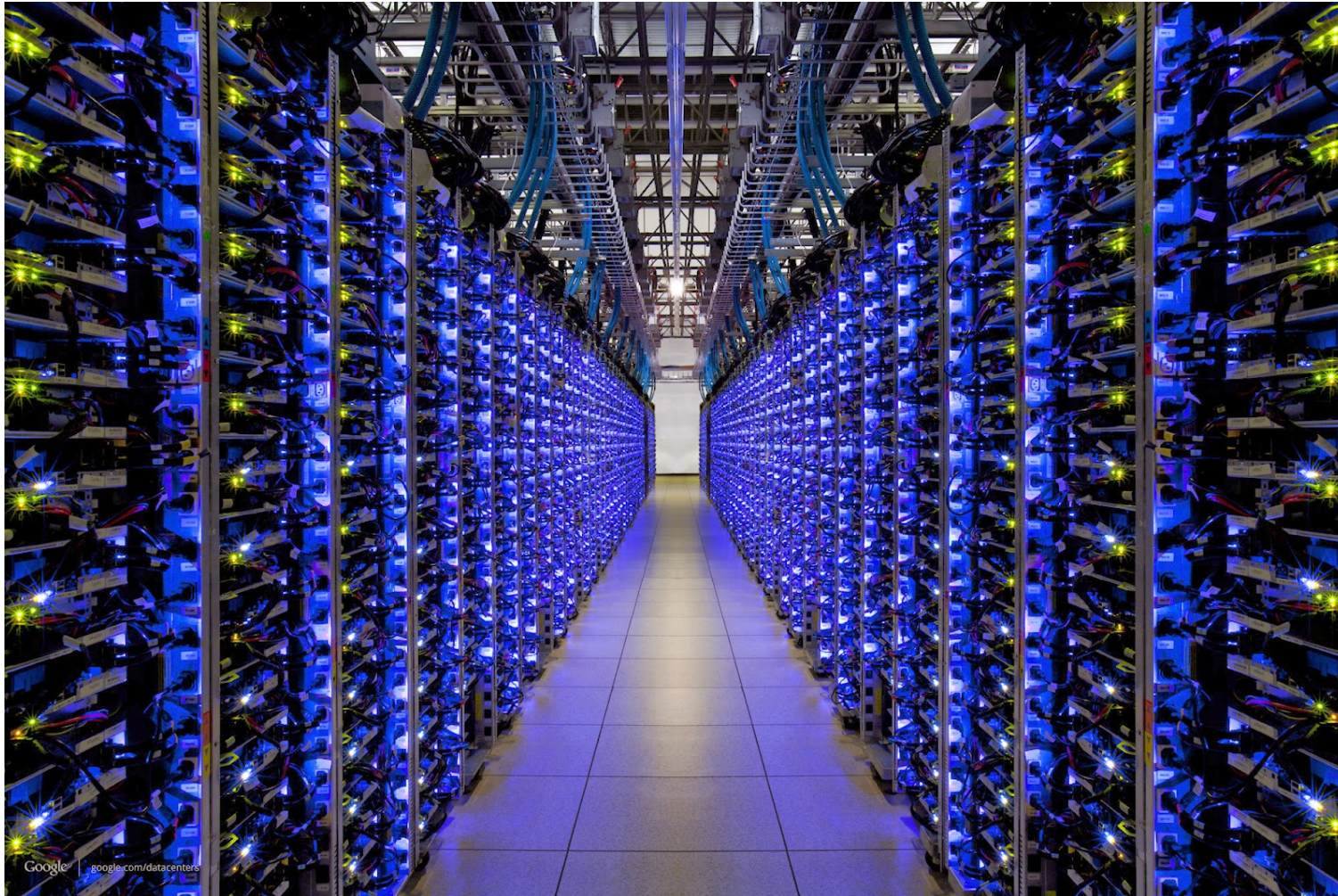
- There are **basic building blocks** and **design principles**



# The Same Applies to Computing Systems

---

- There are **basic building blocks** and **design principles**



# Different Platforms, Different Goals



# Different Platforms, Different Goals

---



# Different Platforms, Different Goals

---



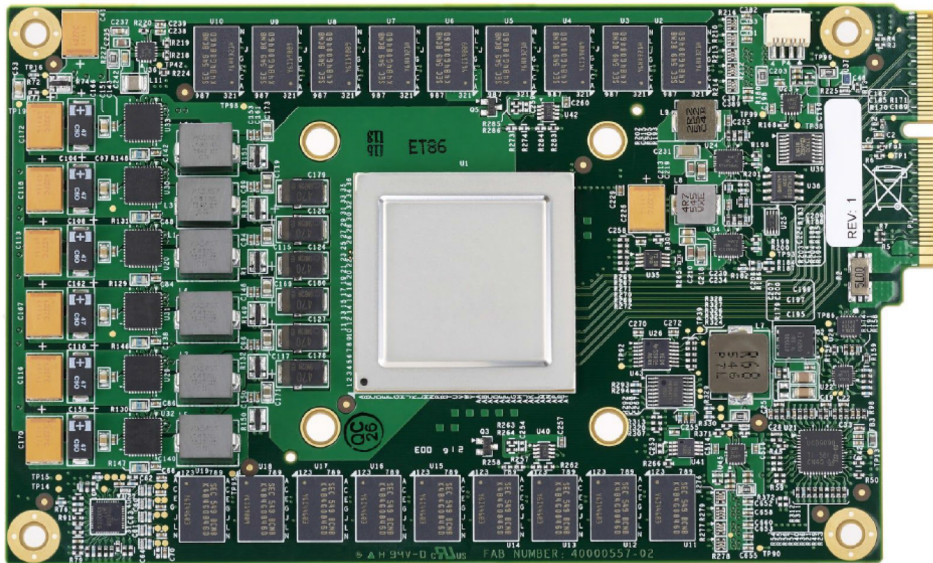
# Different Platforms, Different Goals

---

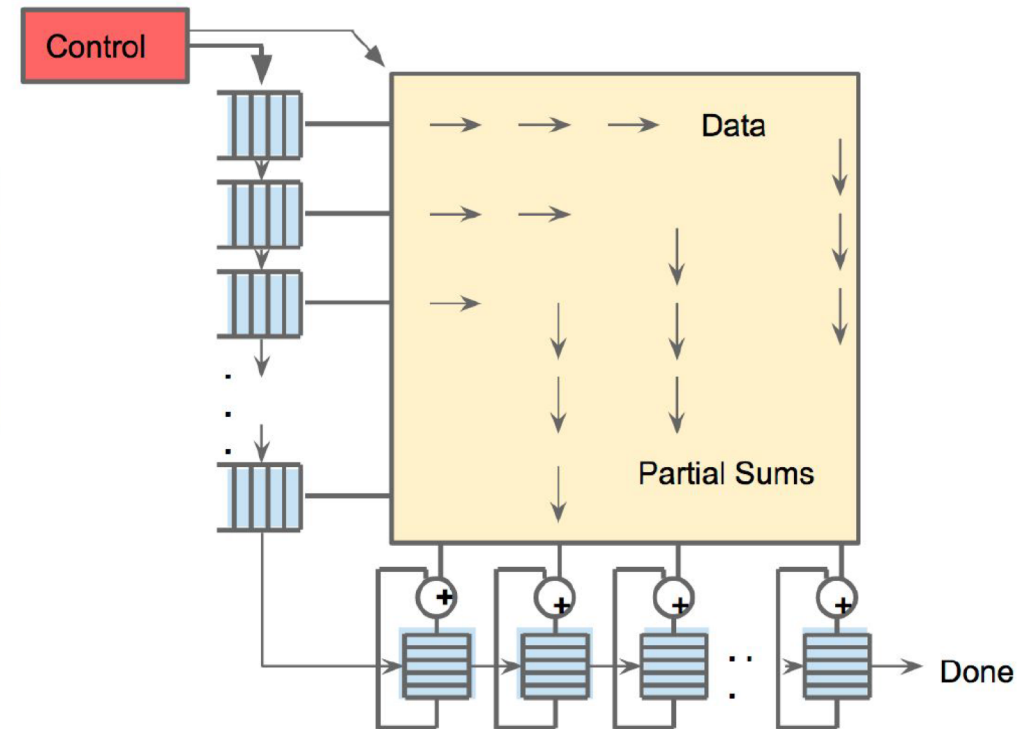


Jack Dongarra

# Different Platforms, Different Goals



**Figure 3.** TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

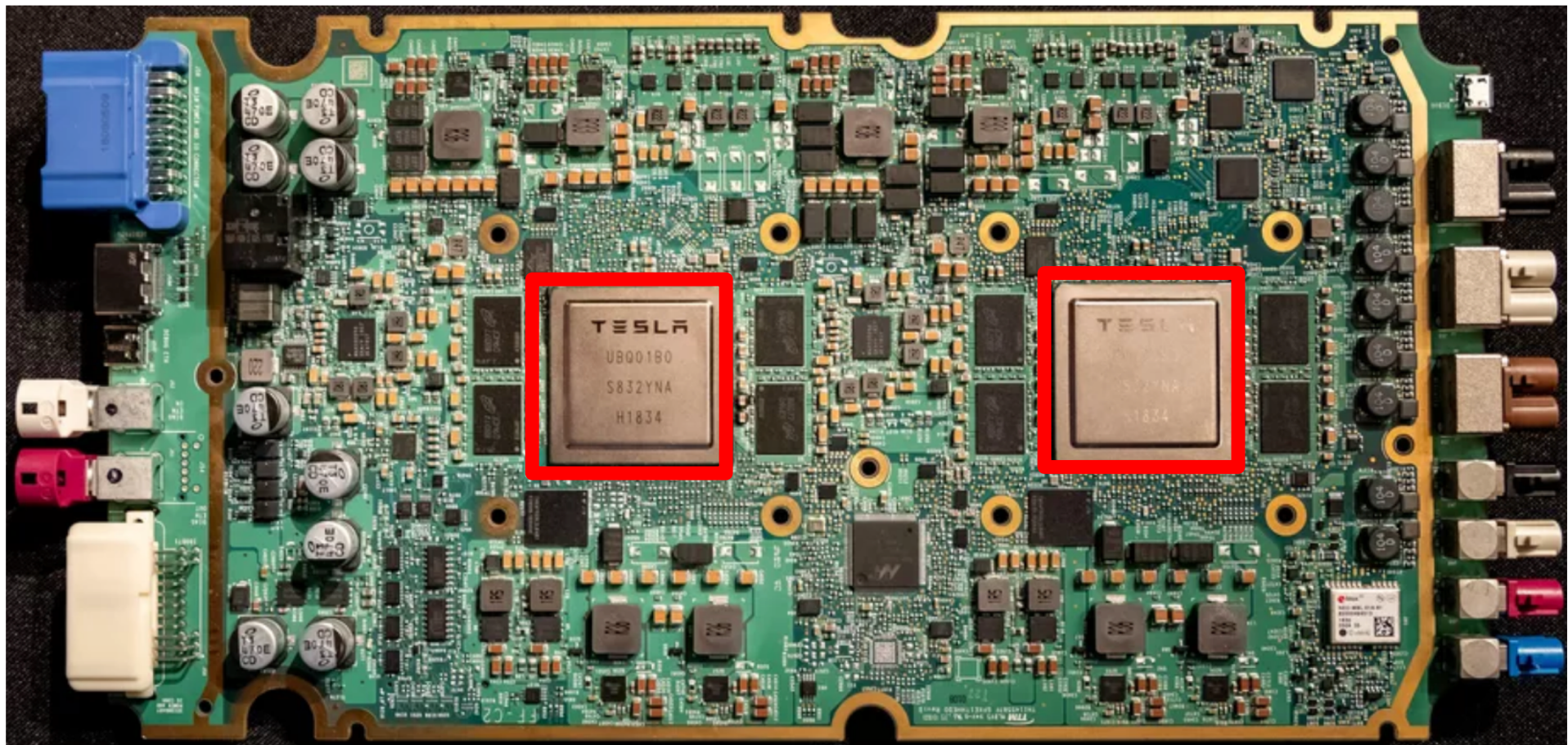


**Figure 4.** Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit”, ISCA 2017.

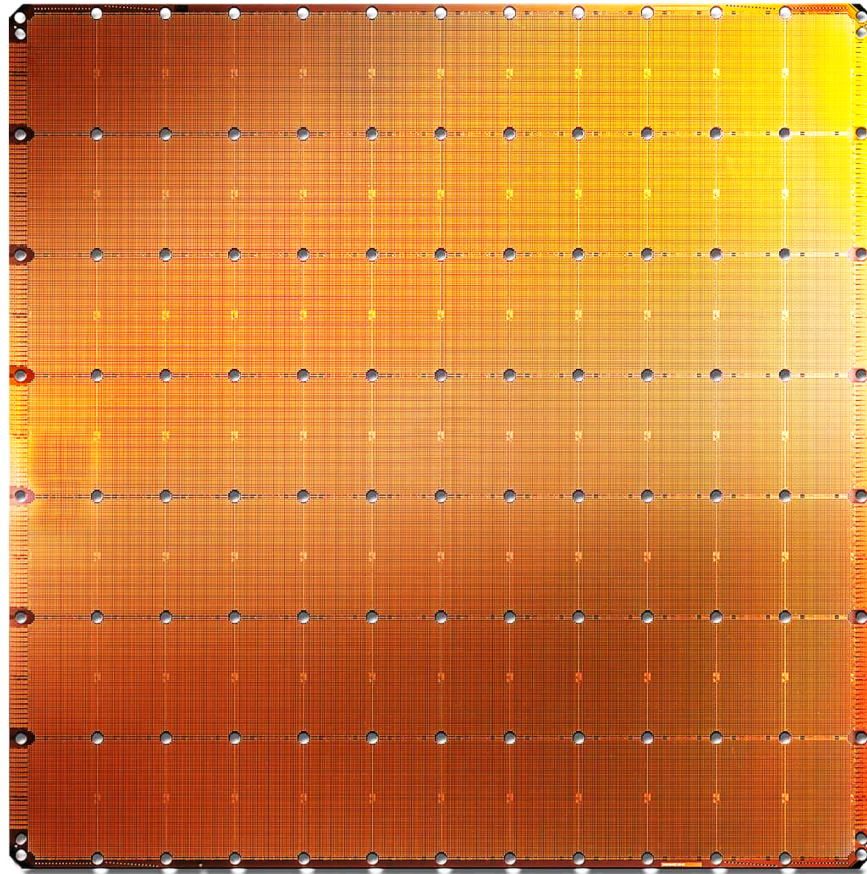
# Different Platforms, Different Goals

- ML accelerator: 260 mm<sup>2</sup>, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.



# Different Platforms, Different Goals

---



## **Cerebras WSE**

1.2 Trillion transistors  
46,225 mm<sup>2</sup>

- The largest ML accelerator chip
- 400,000 cores



## **Largest GPU**

21.1 Billion transistors  
815 mm<sup>2</sup>

NVIDIA TITAN V

<https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning>

<https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/>

# Basic Building Blocks

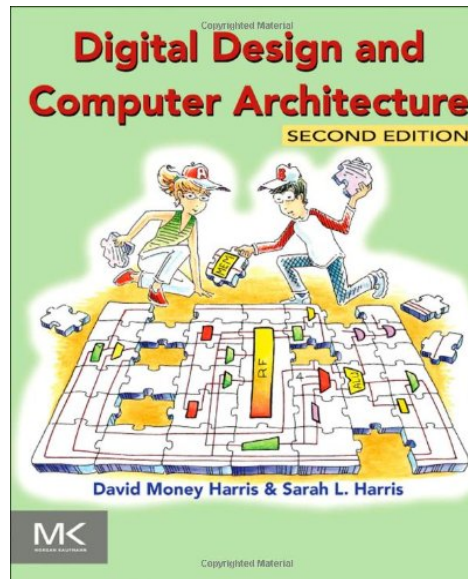
---

- Electrons
- Transistors
- Logic Gates
- Combinational Logic Circuits
- Sequential Logic Circuits
  - Storage Elements and Memory
- ...
- Cores
- Caches
- Interconnect
- Memories
- ...

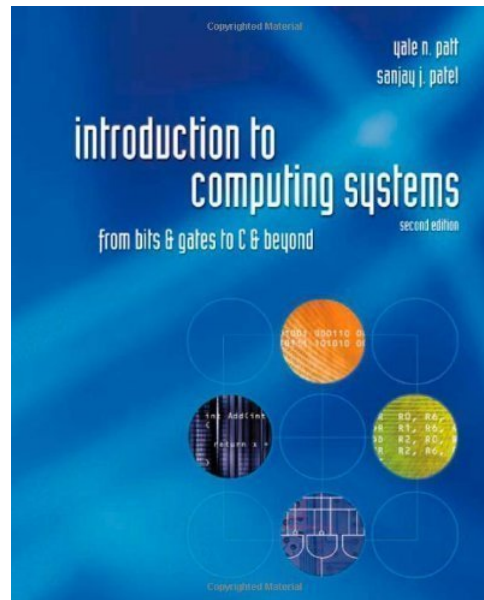
# Reading Assignments for This Week

---

- Chapter 1 in Harris & Harris



- Chapters 1-2 in Patt and Patel



- Supplementary Lecture Slides on Binary Numbers

# Major High-Level Goals of This Course

---

- In Digital Circuits & Computer Architecture
- Understand the basics
- Understand the principles (of design)
- Understand the precedents
- Based on such understanding:
  - learn how a modern computer works underneath
  - evaluate tradeoffs of different designs and ideas
  - implement a principled design (a simple microprocessor)
  - learn to systematically debug increasingly complex systems
  - Hopefully enable you to develop novel, out-of-the-box designs
- The focus is on basics, principles, precedents, and how to use them to create/implement good designs

# Why These Goals?

---

- Because you are here for a Computer Science degree
- **Regardless of your future direction**, learning the principles of digital design & computer architecture will be useful to
  - ❑ design better hardware
  - ❑ design better software
  - ❑ design better systems
  - ❑ make better tradeoffs in design
  - ❑ understand why computers behave the way they do
  - ❑ solve problems better
  - ❑ think “in parallel”
  - ❑ think critically
  - ❑ ...

# Course Info and Logistics

# Course Info: Instructor

---



## ■ Onur Mutlu

- Professor @ ETH Zurich CS, since September 2015 (started May 2016)
- Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-...
- PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
- <https://people.inf.ethz.ch/omutlu/>
- [omutlu@gmail.com](mailto:omutlu@gmail.com) (Best way to reach me)
- Office hours: By appointment (email me)

## ■ Research and Teaching in:

- Computer architecture, computer systems, bioinformatics, hardware security
- Memory and storage systems
- Hardware security
- Fault tolerance
- Hardware/software cooperation
- Genome analysis and application-algorithm-hardware co-design
- ...

# Course Info: Lecturer & PhD Assistants

---

- Head Assistant
  - Dr. Juan Gómez Luna
  
- Vice-Head Assistant
  - Hasan Hassan
  
- Lecturer
  - Dr. Frank Gurkaynak
  
- (Other) Key Assistants and Guest Lecturers
  - Dr. Mohammed Alser
  - Dr. Lois Orosa
  - Dr. Jawad Haj-Yahya
  - Dr. Jisung Park

# Course Info: PhD Assistants

---

- (Other) Key Assistants and Guest Lecturers (cont.)
  - Minesh Patel
  - Giray Yaglikci
  - Can Firtina
  - Geraldo De Oliveira Junior
  - Rahul Bera
  - Konstantinos Kanellopoulos

# Course Info: Student Assistants

---

- Roknoddin Azizibarzoki
- Tim Fischer
- Lukas Gygi
- Leo Horné
- Lara Lazier
- Artur Melo
- Chris Mnuk
- Nathan Neike
- Arpan Prasad
- Nina Richter
- João Dinis Sanches Ferreira
- Taha Shahroodi
- Roberto Starc

# Course Info: Lab Assistants (I)

---

- Tuesday 15-17

- TBD

- Wednesday 15-17

- TBD

# Course Info: Lab Assistants (II)

---

- Friday 8-10

- TBD

- Friday 10-12

- TBD

# If You Need Help

---

- Post your question on Q&A Forum (soon announced)
  - **Preferred** for **technical** questions
  
- Write an e-mail to:
  - [digitaltechnik@lists.inf.ethz.ch](mailto:digitaltechnik@lists.inf.ethz.ch)
  - The instructor and all assistants will receive this e-mail
  
- Come to office hours (CAB H 31.2)
  - Monday 1:30pm-2:30pm
  - Tuesday: 5pm-6pm
  - Wednesday: 10am-11am
    - We might need to change the room due to space limitations. In that case, we will announce it in advance

# Where to Get Up-to-date Course Info?

---

- Website:

- <https://safari.ethz.ch/digitaltechnik/>
- Lecture slides and videos
- Readings
- Lab information
- Course schedule, handouts, FAQs
- Software
- Plus other useful information for the course
- Check frequently for announcements and due dates
- This is your single point of access to all resources

- Your ETH Email

- Lecturers and Teaching Assistants

# Lecture and Lab Times and Policies

---

## ■ Lectures:

- Thursday and Fridays, 13:15-15:00
- HG F7 (F5 overflow)
- Attendance is for your benefit and is therefore important
- Some days, we will have guest lectures and exercise sessions

## ■ Lab sessions:

- See online
- You should definitely attend the lab sessions
  - In-class evaluation (70%) and mandatory lab reports (30%)
- Labs will start on February 28th
- Lab information and handouts are here:
  - <https://safari.ethz.ch/digitaltechnik/spring2020/doku.php?id=labs>

# Lab Organization

---

## ■ Groups

- Choose your preferred group in Moodle

- <https://moodle-app2.let.ethz.ch/mod/choicegroup/view.php?id=412173>

- Due 24.02.2020 at 11:59pm

- Choose your partner

- <https://moodle-app2.let.ethz.ch/mod/feedback/view.php?id=418396>

- Due 24.02.2020 at 11:59pm

## ■ Lab grades from previous years

- <https://moodle-app2.let.ethz.ch/mod/choice/view.php?id=412175>

- Choose among (due 26.02.2020 at 11:59pm):

- 1) I will use my lab grades from previous years, and I won't do the labs this year
- 2) I will use my lab grades from previous years, but I will do the labs this year
- 3) I won't use my lab grades from previous years. I will do the labs this year

# Final Exam

---

- 180-minute written exam
  - Find examination rules in Course Catalogue
  - Also in the first page of previous exams
    - <https://safari.ethz.ch/digitaltechnik/spring2020/doku.php?id=exams>
  - Some exam questions are similar to questions in **Optional HWs**
    - Optional HWs are optional, but **highly recommended**

# Demystifying Mysteries

# Levels of Transformation

“The purpose of computing is [to gain] insight” (*Richard Hamming*)  
*We gain and generate insight by solving problems*  
*How do we ensure problems are solved by electrons?*

## Algorithm

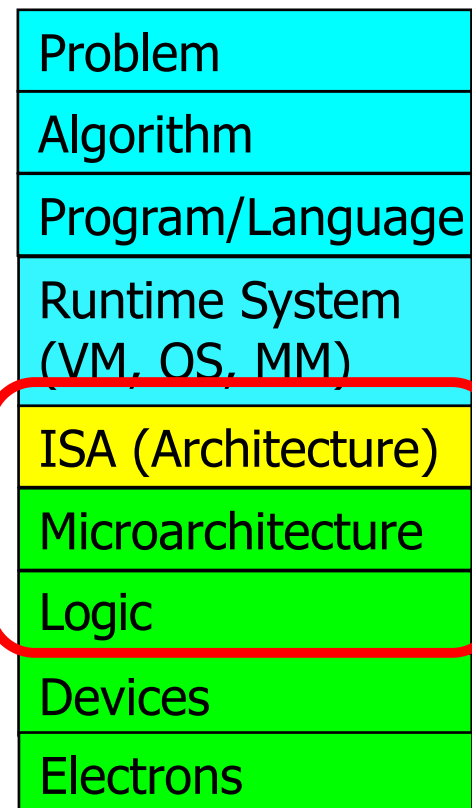
Step-by-step procedure that is **guaranteed to terminate** where **each step is precisely stated** and **can be carried out by a computer**

- **Finiteness**
- **Definiteness**
- **Effective computability**

Many algorithms for the same problem

## Microarchitecture

An implementation of the ISA



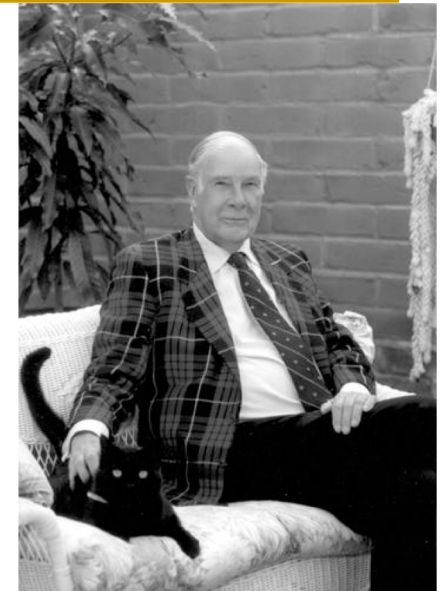
ISA  
(Instruction Set Architecture)

Interface/contract between  
SW and HW.

What the programmer  
assumes hardware will  
satisfy.

## Digital logic circuits

Building blocks of micro-arch (e.g., gates)



# Aside: A Famous Work By Hamming

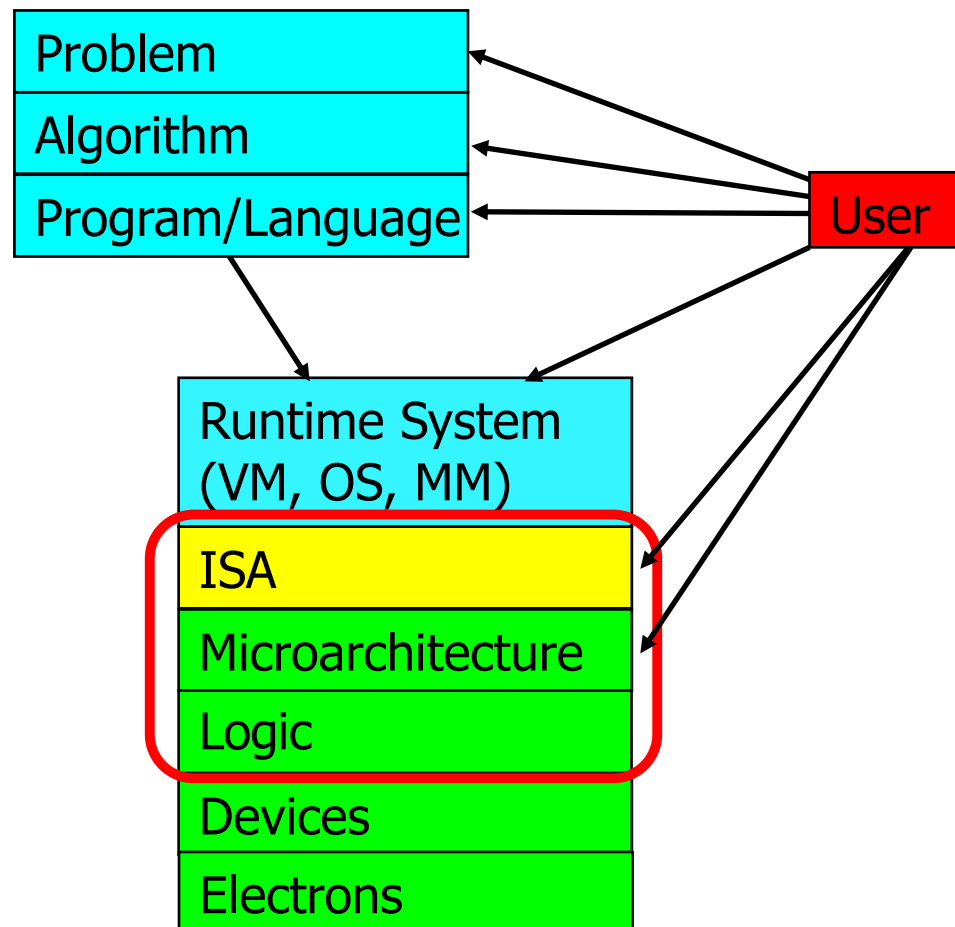
---

- Hamming, “Error Detecting and Error Correcting Codes,” Bell System Technical Journal 1950.
- Introduced the concept of Hamming distance
  - number of locations in which the corresponding symbols of two equal-length strings is different
- Developed a theory of codes used for error detection and correction
- Also see:
  - Hamming, “You and Your Research,” Talk at Bell Labs, 1986.
  - <http://www.cs.virginia.edu/~robins/YouAndYourResearch.html>

# Levels of Transformation, Revisited

---

- A user-centric view: computer designed for users



- The entire stack should be optimized for user

# The Power of Abstraction

---

- Levels of transformation create abstractions
  - Abstraction: A higher level only needs to know about the interface to the lower level, not how the lower level is implemented
  - E.g., high-level language programmer does not really need to know what the ISA is and how a computer executes instructions
- Abstraction improves productivity
  - No need to worry about decisions made in underlying levels
  - E.g., programming in Java vs. C vs. assembly vs. binary vs. by specifying control signals of each transistor every cycle
- Then, why would you want to know what goes on underneath or above?

# Crossing the Abstraction Layers

---

- As long as everything goes well, not knowing what happens underneath (or above) is not a problem.
- What if
  - The program you wrote is running slow?
  - The program you wrote does not run correctly?
  - The program you wrote consumes too much energy?
  - Your system just shut down and you have no idea why?
  - Someone just compromised your system and you have no idea how?
- What if
  - The hardware you designed is too hard to program?
  - The hardware you designed is too slow because it does not provide the right primitives to the software?
- What if
  - You want to design a much more efficient and higher performance system?

# Crossing the Abstraction Layers

---

- Two goals of this course (especially the second half) are
  - to understand how a processor works underneath the software layer and how decisions made in hardware affect the software/programmer
  - to enable you to be comfortable in making design and optimization decisions that cross the boundaries of different layers and system components

# Some Example “Mysteries”

# Digital Design & Computer Arch.

## Lecture 1: Introduction and Basics

Prof. Onur Mutlu

ETH Zürich

Spring 2020

20 February 2020

We Did Not Cover the  
Following Slides in Lecture 1